Azure Data Lake Storage Gen 2 -säilytyskustannusten hallinta palvelutason (access tier) optimoinnin avulla

Nina Eerikäinen

Haaga-Helia ammattikorkeakoulu
Tradenomi, tietojenkäsittely
Amk-opinnäytetyö
kevät 2024

Tiivistelmä

Tekijä(t)
Eerikäinen Nina

Tutkinto
Tradenomi, tietojenkäsittely

Raportin/Opinnäytetyön nimi
Azure Data Lake Storage Gen 2 -säilytyskustannusten hallinta palvelutason (access tier) optimoinnin avulla

Sivu- ja liitesivumäärä
39 + 1

Opinnäytetyö toteutettiin laadullisena tutkimuksena, ja se on yhdistelmä kirjallisuuskatsausta ja itse luotuihin käyttötapauksiin pohjautuvaa empiiristä tutkimusta. Työ käsitteli tiedon säilytyskustannuksia data-analytiikan käyttöön tarkoitetussa Azure Data Lake Storage Gen 2 -tietoaltaassa ja siinä selvitettiin säilytyskustannusten alentamisen mahdollisuuksia etsimällä esimerkinomaisille käyttötapauksille sopivin tiedon säilyttämisen palvelutaso (access tier). Tietoperustassa käytettiin pääosin Microsoftin tuottamaa materiaalia, jota oli luettavissa heidän verkkosivuiltaan tammi-toukokuun 2024 välisenä aikana. Kirjallisuudesta hyödynnettiin myös Azure-sertifiointeihin valmistavaa kurssimateriaalia, erilaisia opaskirjoja, blogitekstejä ja artikkeleita.

Kustannusten arviointi perustui Microsoftin kustannuslaskuriin syötettyihin säilytyskapasiteetin kokoa, tietoihin kohdistuneita luku-operaatioita ja niiden sekä tiedostojen keskimääräistä kokoa kuvaaviin lukuihin. Nämä luvut saatiin selville lokitiedoista Azuren Log Analytics -palvelussa ja tietoaltaaseen talletettujen tietojen tarkastelun mahdollistavassa Azure Storage Explorer -sovelluksessa. Opinnäytetyössä osoitettiin lokitietojen merkitys datan tuntemisen keskeisenä tekijänä, ja mahdollisuus ennustaa tulevaa datan käyttöä tutkimalla menneitä lokimerkintöjä. Tiedostokokojen ja säilytyskapasiteetin arvioinnin yhteydessä esiteltiin binääristen ja metristen tallennustilaa koskevien yksiköiden – kuten tebitavujen ja teratavujen – väliset erot.

Tiedon säilyttämisen palvelutasoja on Azure Data Lake Storage Gen 2 -tietoaltaassa neljä: kuuma, viileä, kylmä ja arkistointiin tarkoitettu palvelutaso. Periaate palvelutasojen hinnoittelussa on se, että kuumalla palvelutasolla säilytystila on kallista ja tiedon käyttö halpaa, kun taas tiedon käytön tuottamat kustannukset kasvavat viileämmille – säilytystilaltaan halvemmille – palvelutasoille siirryttäessä. Käyttötapausten avulla opinnäytetyössä todistettiin tämän periaatteen paikkansapitävyys, ja tarjottiin kuhunkin tapaukseen soveltuvia vinkkejä kustannusten alentamiseen. Esimerkkinä kustannusten alentamiseen liittyvästä ratkaisusta esitettiin Microsoftin tarjoama Azure Lifecycle Management -palvelu, jolla käyttämättömän datan saa automaattisesti siirrettyä edullisemmalle palvelutasolle.

Tietoperustassa ja käyttötapausten läpikäynnin yhteydessä tuotiin esille myös joitakin rajoituksia ja huomioon otettavia seikkoja liittyen tiedon käyttöön ja sen saatavilla olemiseen, minimisäilytysaikaan kullakin palvelutasolla ja tiedon siirrosta palvelutasolta toiselle aiheutuviin kustannuksiin.

Johtopäätöksinä opinnäytetyössä esitetään, että säilytyksen palvelutason optimointi vaatii etukäteissuunnittelua, datan tuntemista ja sen luokittelua kullekin palvelutasolle sopivimpiin kokonaisuuksiin. Huomiota kiinnitetään myös tämän tavoitetilan saavuttamisen käytännön haasteisiin, sekä pilvioperaattorin hinnoittelun monitahoisuuteen.

Asiasanat
Data, säilyttäminen, kustannukset

Sisällys

1 Johdanto

Opinnäytetyöni aihe on Azure Data Lake Storage Gen 2 -tietoaltaan säilytyskustannusten hallinta tiedon säilytykseen liittyvän palvelutason (*access tier*) avulla. Päädyin aiheeseen uteliaisuudesta selvittää tämän Microsoft Azure -palvelun hinnoittelua käytännön tapauksissa ja toisaalta mielenkiinnosta selvittää tiedon säilytyksen kustannusten alentamisen vaihtoehtoja. Olen myös työssäni pohtinut, voisiko palvelutason muuttaminen olla yksi ratkaisu tilanteisiin, joissa harvoin käytettäviä tietoja ei haluta kokonaan poistaa. Lisäksi minua kiinnostaa selvittää Azuren lokeja monitoroimalla, miten tietoaltaassa säilytettävää dataa todella käytetään.

Microsoft Azure Data Lake Storage Gen 2 on yleisesti käytössä oleva data-analytiikkaan liittyvän tiedon säilyttämisen palvelu. Tämän kaltaisille analytiikan tietoaltaille on tyypillistä, että niihin kerätään ja niissä säilytetään suuria määriä dataa. Pilvikustannusten hallinta mietityttää monissa organisaatioissa, joissa siirtymä on-premises -ratkaisuista pilveen on tapahtunut hiljattain tai joissa vasta pohditaan pilvisiirtymää. Hinnoittelun periaatteista ja kustannuksiin vaikuttavista tekijöistä voi olla vaikeaa saada otetta. Tiedon säilytyksen palvelutasojen vaihtoehtoja kartoittaessa olisi niin ikään hyvä saada listatuksi niiden edut ja haitat niin kustannusten kuin esimerkiksi tiedonhakuviiveen (*latency*) suhteen. Tutkimukseni voi olla hyödyllinen tiedon säilytyksen kustannuksia seuraaville tahoille erityisesti arkkitehtien, datainsinöörien tai kustannuksia seuraavien johtajien tai tuoteomistajien rooleissa.

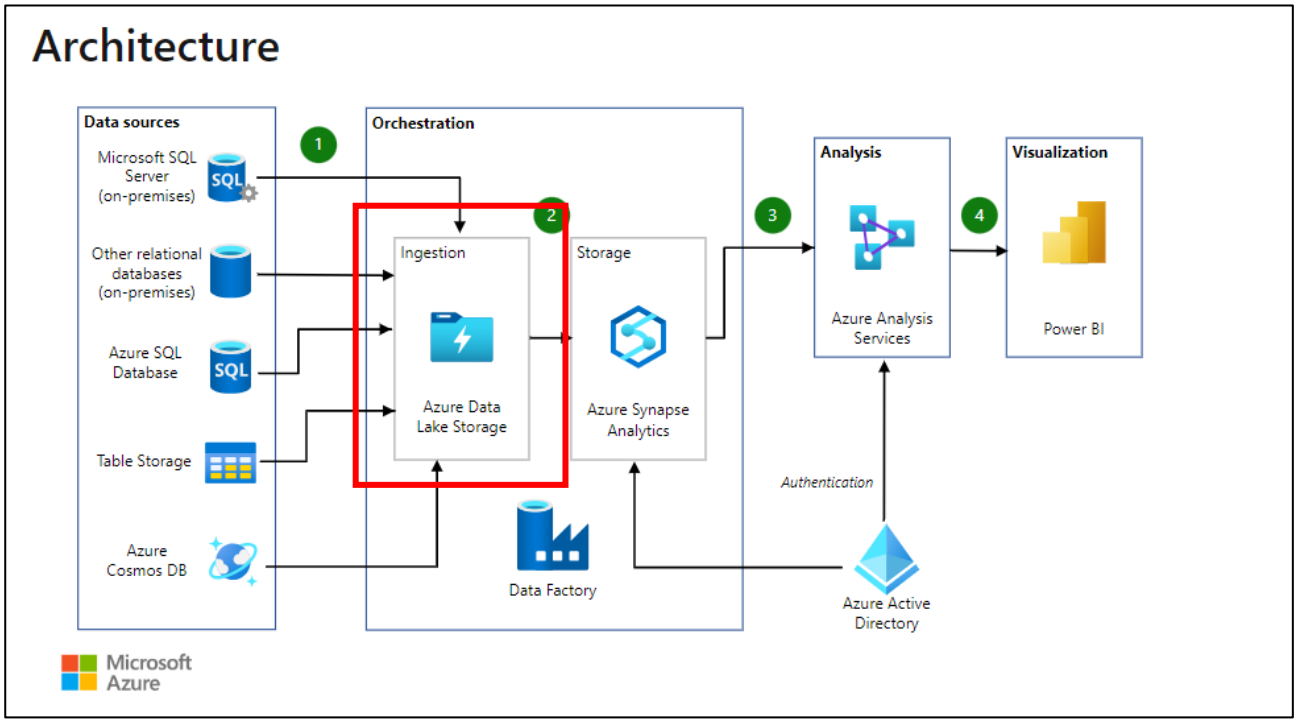
Opinnäytetyössäni etsin kustannustehokkainta ratkaisua kolmenlaiseen tilanteeseen, joissa tiedon säilytykseen kohdistuu erilaisia tarpeita (käyttötapaukset). Taustoitan aihettani esittelemällä tässä luvussa 1 dataputki-arkkitehtuurin kautta tietoaltaan ideaa ja keskeisiä käsitteitä. Luvussa 2 kerron tärkeimmistä lähteistäni, aiemmasta tutkimuksesta ja Microsoft Azure Data Lake Storage Gen 2 -palvelun ja sen tiedon säilytyksen hinnoitteluperiaatteista. Luvussa 3 selvitän, miten Microsoftin kustannuslaskuria varten tarvittavat tiedostokoot ja tietoihin kohdistuvaa käyttöä kuvaavat luvut saadaan selville käyttämällä Azure Storage Explorer- ja Microsoft Log Analytics -palveluita. Luvussa 4 esitän kolmen käyttötapauksen kautta, miten säilytyksen palvelutasoon liittyviä kustannuksia voidaan optimoida. Luvussa 5 on tulosten analysoinnin, johtopäätösten ja oman oppimisen arvioinnin aika.

1.1 Aiheen tausta, rajaukset ja tutkimuksen konteksti

Dataputken avulla voidaan kuvata datan kulkua sen saapumisesta tietoaltaaseen aina siihen saakka, kun se päätyy datatuotteeksi esimerkiksi analytiikan sovelluksen tai raportoinnin käyttöön.

Dataputki koostuu erilaisista komponenteista, jotka liittyvät datan lukemiseen, muokkaukseen, jalostamiseen, analysointiin, tallennukseen ja sen ”loppukäyttöön”. Yksittäisillä komponenteilla voi olla eri kehittäjätahoja ja erilainen elinkaari. (Tietoevry 5.5.2020).

Microsoftin viitearkkitehtuuri-esimerkissä on erotettavissa dataputken osina tietolähteet (kuvassa merkittynä numerolla 1), datan lataaminen ensin Azuren Data Lake Storage -palveluun ja siitä edelleen Azure Synapseen (kohta 2), tiedon käsittely analytiikan menetelmin (kohta 3) ja datan visualisointi PowerBI:n avulla (kohta 4). (Microsoft, s.a.a). Käsittelen opinnäytetyössäni Azuren Data Lake Storage Gen 2 -tietoallasta, joka on korostettu kuvassa punaisella. Käytän palvelusta jatkossa myös lyhennettä ADLS Gen 2. Tietoallas voi olla datan ainoa säilyttävä komponentti, tai kuten oheisessa kuvassa, data voi kulkea sen läpi edelleen säilytettäväksi toisessa komponentissa.



Kuva 1. Microsoftin dataputki esitettynä viitearkkitehtuuri -kuvana (Microsoft s.a. a)

Analytiikan tarpeisiin kerätään dataa niin organisaation sisäisistä operatiivisista tietojärjestelmistä, kuin ulkoisista lähteistäkin. Data-analytiikka mahdollistaa päätöksenteon perustelemisen, liiketoiminnan kehittämisen, ennusteiden laatimisen ja trendien seuraamisen. (Törmänen 2017, 4; 108-109).

Azure Data Lake Storage Gen2:ssa säilytettävien tietojen käsittely tapahtuu esimerkiksi Databricksin tai Synapse Analyticsin kaltaisella data-alustalla. Kun ADLS Gen2 -tietoaltaassa säilytettäviin tietoihin kohdistuu käyttöä, siitä jää lokimerkintä Microsoftin Log Analytics -palveluun, mikäli tietoa

käsittelevä applikaatio on liitetty lokituksen piiriin.

1.2 Tavoitteet ja tutkimuskysymykset

Tavoitteenani on selvittää kolmea käyttötapausta vertaillen, minkälaisissa tilanteissa tiedon säilyttäminen viileällä, kylmällä tai arkistointiin tarkoitetulla palvelutasolla tulisi edullisemmaksi kuin sen pitäminen kuumalla palvelutasolla. Pyrin pohjaamaan käyttötapaukset mahdollisuuksien mukaan realistiseen tietoaltaan käytön monitorointiin. Tutkin myös erilaisten säilyttämiseen liittyvien palvelutasojen vaikutuksia tietojen hakemisen viiveeseen ja selvitän, mitä muita seikkoja on mahdollisesti otettava huomioon palvelutasojen muutoksissa.

Microsoft Azure -palveluiden lokeja voidaan tutkia Azure Portal -käyttäjäportaalin Log Analytics -lokitietosovelluksen kautta. Lokeista voidaan hakea esille kustannusarvioiden pohjaksi tarvittavia luku- ja kirjoitus -operaatioiden määriä ja myös operaatioiden suorittajatahoja (applikaatiot tai yksittäiset käyttäjät). Tutkimalla luku-operaatioiden määriä ja toistumissyklejä esimerkiksi samaan tiedostokansioon tallennetuista tiedoista etsin opinnäytetyöhöni tapauksia, joille laskea kustannukset suhteessa viileälle, kylmälle tai arkistointiin tarkoitetulle palvelutasolle siirtämiseen ja siellä säilyttämiseen ja kuumalla palvelutasolla pitämiseen.

Arvioin kustannuksia tiedoille, joita jo säilytetään ADLS Gen 2 -palvelussa. Oletuksena tiedot on tallennettu kuumalle palvelutasolle. Kustannusten laskennassa tarvittavat tiedostojen koot saan selville Azure Storage Explorer -sovelluksen kautta. Tämä tieto on ilmoitettu gibitavuina ja mebitavuina.

Jätän tutkimukseni ulkopuolelle muun tyyppiset kustannukset kuin säilyttämisestä, tiedon siirtämisestä eri palvelutasojen välillä ja luku-operaatioista syntyvät. Kustannusmuuttujien monimutkaistaminen tekisi erilaisten skenaarioitten takia tutkimuksesta liian rönsyilevän. ADLS Gen 2 -palvelussa pakollisia kustannuksia syntyy tiedon säilytystilasta, jokaisesta kerrasta kun tietoa luetaan tai siihen kirjoitetaan (luku- ja kirjoitusoperaatiot) ja tiedon siirrosta. Valinnaisia tai täydentäviä kustannuksia syntyy esimerkiksi metatiedoista ja kyselyiden optimoinnista. (Microsoft 2023a). Lähtökohtana opinnäytetyössäni on, että tietoa siirretään kustannussäästöä hakien lämpimämmältä palvelutasolta viileämmälle, ei toisinpäin. Hintatiedot ovat suuntaa-antavia, ja ne perustuvat Microsoftin huhtikuussa 2024 ilmoittamiin summiin.

Teen lokitietojen analysointia ainoastaan Azure Portalin kautta KQL-kieltä käyttäen. Pyrin näin ennustamaan datan tulevaa käyttöä katsomalla menneeseen. Microsoftilla on Cost Management -palvelu asiakasportaalissaan, mutta en käytä sitä tämän opinnäytetyön kustannusarvioissa. Cost Management on tarkoitettu Azure-palveluihin tai muiden siihen kytkettyihin palveluihin liittyvien

kustannusten analysointiin, visualisointiin ja kustannusten ennustamiseen. Sen Advisor -toiminnon ehdotukset voivat liittyä myös säilytyskustannusten alentamiseen.

Microsoft suosittelee kustannusten hallintaan tietoaltaan sisällön säännöllistä inventointia ja analysointia. Tässä työssä esitettyjen keinojen sijasta näkymiä tietoaltaan sisältöön voidaan toteuttaa myös esimerkiksi Databricks-alustan kautta Python-koodia käyttäen. (Microsoft 2023a).

Azuren pay-as-you-go -tilauksen laskutus on kuukausittaista, ja teen kustannusarvioni niin ikään kuukausitasolla. Käyttötapauksessa 2 lasken myös oletetut vuosikustannukset. Käyttötapauksissani data on tiedostoina suurissa tietokannoissa ja se on muodoltaan rakenteista, tietokantamuotoista tietoa - ei esimerkiksi kuvaa, ääntä, tai "puolirakenteista" dataa.

Tutkimuskysymykseni ovat:

> *Miten saadaan selville Microsoft Azure Data Lake Storage Gen 2 -palvelun säilytyskustannusten arvioinnissa tarvittavat lukumäärätiedot luku-operaatioille ja tiedostojen koolle?*
>
> *Minkälaisessa tilanteessa datan siirtäminen kuumalta palvelutasolta viileämmälle palvelutasolle tuottaa kustannussäästöä?*
>
> *Minkälaisia rajoituksia viileämmällä palvelutasolla säilyttämiseen liittyy ja mitä tulee ottaa huomioon siirrettäessä dataa palvelutasojen välillä?*
>
> *Miten palvelutasoon liittyvää kustannusoptimointia voidaan käytännössä edistää ja toteuttaa?*

Tutkimukseni vastuullisuusnäkökulma syntyy kustannusten säästämisestä. Yhteiskunnassa on taloustilanteen vuoksi painetta karsia turhia menoja. Lisäksi kalliimman palvelutason vaatima levytila on konkreettisesti enemmän energiaa vievää, kuin ”viileämpi” levytila. Optimaalisen palvelutason valinta vaikuttaa tarvittavaan levytilakapasiteettiin ja säästää energiaa pilvipalveluiden tarjoajan konesaleissa. Vastuullisen henkilötietojen käsittelyn näkökulmasta huomautan, että tietojen saattaminen viileämmälle palvelutasolle ei tarkoita niiden asianmukaista poistamista, vaan tämä on suunniteltava erikseen.

1.3 Keskeiset käsitteet

Azure Data Lake Storage Gen 2, ADLS Gen2

Suurten datamäärien keräämisen, säilyttämisen ja analysoinnin mahdollistava tietoallas. Yhdistää rakenteisen datan keräämisen ja tallentamisen Microsoftin Blob Storagen (objektien ja ei-rakenteisen tiedon hallinnan) ominaisuuksien kanssa, jolloin esimerkiksi tietoturvaa tai skaalautuvuutta voidaan hallita tiedostotasolla. (Borgini 2023).

Azure Storage Explorer

Työpöytäsovellus, jonka avulla ADLS Gen 2 -palveluun tallennettuja tietoja voidaan tarkastella. (Microsoft 2023b).

Azure Storage Lifecycle Management

Työkalu, jolla Blob Storage -pohjaiseen tietoaltaaseen voidaan luoda tietojen säilyttämistä koskevat säännöt. Säännöistä riippuen dataa voidaan siirtää kuumemmalta palvelutasolta viileämmälle tai toisinpäin perustuen siihen, milloin dataa on viimeksi käytetty. (Microsoft 2023c).

Blob

Blob on lyhenne sanoista *Binary Large Object* (Pure Storage 2024). Azure Data Lake Storage Gen 2 pohjautuu Blob storage -rakenteeseen. (Microsoft 2023d). Blobeja voidaan ajatella synonyymina tiedostoille. ADLS Gen2 -palvelussa tiedostoista tulee blobeja, mutta tämä ei vaikuta niiden käyttöön. (Microsoft 2023e).

Sana ”blob” suomentuu binäärimuotoiseksi, suureksi objektiksi. Säilytyksen yhteydessä käsite ”objekti” viittaa tiedostomaiseen olemukseen, ja mihin tahansa tallennusformaattiin. Objekti voi siis olla tekstiä, kuvia, liikkuvaa kuvaa tai audiota. Objekteina tallennettu tieto on hyvin skaalautuvaa. (Reis & Housley 2022, 209).

Data-analytiikka

Laadullisia ja määrällisiä tekniikoita ja prosesseja, joiden avulla organisaatiot pyrkivät tehostamaan tuottavuutta ja kasvattamaan liiketoimin-

taansa (Törmänen 2017, 179). Analytiikka voidaan jakaa myös kuvailevaan ja ennustavaan (Ari Hovi Oy 2023). Yksi kuvailevan analytiikan muoto on *Business Intelligence* (Salesforce 2024).

Kuuma, viileä, kylmä ja arkistointiin tarkoitettu palvelutaso

Microsoftin Azure Data Lake Storage Gen 2 -tietoaltaassa mahdolliset säilytykseen liittyvät palvelutasot ovat *Hot tier*, *Cool tier*, *Cold tier* ja *Archive tier*. Palvelutaso vaikuttaa tiedon säilytyksen kustannuksiin pilvessä. Kuuma palvelutaso on kallein tiedon säilyttämiseen, mutta viileän, kylmän ja arkistointiin tarkoitetun palvelutason kustannuksiin vaikuttaa tiedon käyttö. (Microsoft 2023d).

Luku-, kirjoitus- ja laskenta -operaatiot

Tietokoneen luku-, kirjoitus- ja laskentatoiminnot (*read*, *write*, *compute)*. Luku-operaatiossa tietokone palauttaa sen muistiin tallennetun tiedon, kun taas kirjoitus-operaatiossa muistiin tallennetaan uusi arvo. (University of Babylon 2018, 1). Esimerkiksi SQL-kysely muodostaa luku-operaation (Dremio 2024). Näiden lisäksi pilvipalveluissa puhutaan laskenta -operaatioista tarkoittaen tiedon prosessointia, muistin käyttöä, verkkopalveluita ja tallennusta, joita tarvitaan ohjelman onnistuneeseen suorittamiseen. (Amazon Web Services 2024).

Microsoft Log Analytics

Microsoft Azure -asiakasportaalin kautta käytettävä työkalu lokitietojen kyselyyn ja analysointiin. Kyselyissä käytetään Kusto Query Language (KQL) -kieltä. (Microsoft 2023f).

Säiliö, *container*

Säiliö voidaan nähdä synonyymina tiedostojärjestelmälle. (Microsoft 2023g). Blobit on tallennettava säiliöihin. Säiliö järjestää blobit hakemistomaiseen rakenteeseen, kuten tiedostojärjestelmissä. Tallennustilillä voi olla rajaton määrä säiliöitä ja säiliöihin voi tallentaa rajattomasti blobeja. (Microsoft 2023h).

Tallennustili, *storage account*

Tallennustili, jonka alle organisaation Azure Data Lake Storage Gen 2 ja muut tiedon säilyttämiseen ja tallentamiseen liittyvät Azure -palvelut tietoineen kuuluvat. Tallennustili tarjoaa yksilöllisen HTTP-nimiavaruuden kaikelle tietoaltaaseen talletetulle tiedolle. Tallennustilin avaaminen tapahtuu Microsoft Azuren asiakasportaalissa. (Microsoft 2023i).

Tietoallas, *data lake*

Tietoallas on keskitetty tietojen säilytyspaikka sekä rakenteiselle, että ei-rakenteiselle tiedolle. Tiedon muotoa ei ole tarpeen muuttaa, vaan se saa olla tietoaltaassa ”raakana”, sellaisena kuin se on. Tietoaltaassa tieto on nopeasti ja helposti saavutettavissa sen analysointia varten. (Microsoft 2023g).

Tietoallas konseptina yleistyi *big data*:n eli suurten datamassojen keräämisen myötä kuvaamaan useasta eri lähteestä kerättävää, ”raakaa” dataa, joka ei välttämättä muodosta perinteistä tietokantarakennetta vaan vertautuu paremmin tiedostojärjestelmään, kuten tietokoneen verkkolevyyn. (Ari Hovi Oy 1.12.2020). Viime vuosina tietoaltaiden ja enemmän perinteisiin relaatiotietokantarakenteisiin perustuvien tietovarastojen rinnalle on noussut *data lakehouse* -ratkaisuja, joissa eri datan käyttäjäryhmien tarpeet tulevat huomioiduksi ja lähdedatan tiedostopohjaisuus yhdistyy perinteiseksi relaatiokannaksi. (Ari Hovi Oy 1.12.2020, 2.11.2021).

2 Tutkimusmenetelmän, lähteiden ja Azuren kustannuslaskurin esittely

Tutkimukseni kohteena on Microsoftin ADLS Gen 2 -tuotteella toteutettu tietoallas ja siellä säilytettävä data. Lisäksi kerron, miten löydetään hinnoittelun pohjaksi tarvittavat tiedostokoot Azure Storage Explorerin avulla ja toisaalta miten Azure Log Analytics -palvelun kautta hallinnoitavia lokitietoja voi käyttää dataan kohdistuvien luku-operaatioiden etsimiseen. Näiden operaatioiden määrällä on merkitystä vertailtaessa eri palvelutasojen hintaa ja käyttökohteita.

Tutkimus on laadullinen ja yhdistelmä kirjallisuuskatsausta sekä käytännön kokeilua (ns. empiirinen osa). Kirjallisuuskatsausten tyypeistä menetelmä on lähimpänä systemaattista kirjallisuuskatsausta. Systemaattisessa kirjallisuuskatsauksessa keskeistä on kontekstin muodostaminen tutkittavalle aiheelle monipuolista lähdeaineistoa tulkitsemalla. Olennaista on löytää merkityksellisiä, mielenkiintoa herättäviä ja toisaalta myös keskenään loogisia lähteitä. Systemaattinen kirjallisuuskatsaus pyrkii huolellisuuteen ja perusteellisuuteen. (Mannila 11.2.2021). Kirjallisuuskatsaukseen liittyy analyyttinen ote, eikä sen ole tarkoitus olla vain lähdeaineiston luettelo. Tarkoitus on arvioida lähteitä myös kriittisesti. (Salminen 2011, 4). Lopputuloksena kirjallisuuskatsauksesta syntyy tuloksista tehtävä synteesi (Salminen 2011, 11), jonka teen kuvailevana opinnäytetyöni pohdintaosuudessa.

Kirjallisuuskatsauksen luonteen mukaisesti käytin lähteiden etsintään varsin paljon aikaa. Etsin lähteitä Googlen ja Haaga-Helian Finna -palvelun kautta. Seuloin lähteistä käytettäviksi vain tuoreimmat ja mahdollisimman täsmällisesti aiheeseen liittyvät. Vanhin varsinaiseen tutkimusaiheeseeni liittyvä lähde on vuodelta 2017. Tutkimuksen tietoperustan keräämiseen olen käyttänyt kirjallisuuden lisäksi pienissä yksityiskohdissa Microsoftin asiantuntijoilta sähköpostitse saamiani vinkkejä.

Käytännönläheisten esimerkkitapausten osuus opinnäytteessäni ei täysin istu kirjallisuuskatsauksen määritelmään. Käytin Chat GPT:n apua tämänkaltaisen tutkimussuuntauksen nimen ja määritelmän etsintään. Itse keksittyihin tai kokemuspohjaisiin käyttötapauksiin perustuvalle tutkimussuuntaukselle ei ole vakiintunutta nimeä, mutta Chat GPT tarjoaa tällaiselle lähestymistavalle termiä "*case-based research*". Kokemuspohjaiset käyttötapaukset perustuvat esimerkkeihin, jotka voisivat olla totta tietoallasta käyttävälle, suurelle organisaatiolle. Haluan konkretisoida Microsoftin hinnoittelumallia todentuntuisilla tietomassoilla, joiden säilyttäminen sopivalla palvelutasolla auttaa säästämään kustannuksia.

Rakennan teoriamaisen osuuden (tietoperusta) Microsoftin verkkomateriaalien ja muun tarkasti seulomani lähdeaineiston varaan ja testaan empiirisessä osassa käytännön tapauksin, miten teoriassa esitetyt asiat ovat todennettavissa. Lopussa teen johtopäätöksiä, yhteenvetoja ja analysoin oppimaani.

2.1 Lähteet ja aiempi tutkimus

Keskeisen tietoperustan opinnäytetyössäni muodostaa Microsoftin Azure Learn -verkkomateriaali ja toisaalta Microsoft Azure -asiakassivusto alasivuineen. Näiden verkkosivustojen kautta löytyvät linkit Azure-palveluiden esittelyihin, joita käytän myös käsitteiden määrittelemisessä. Lisäksi niiden kautta pääsee tutustumaan hinnoitteluperiaatteisiin ja esimerkiksi lokien analysointiin käytettävää Kusto Query Language -kyselykieltä koskeviin oppimismateriaaleihin. Käytän englanninkielistä versiota Microsoftin lähdeaineistosta.

Microsoftin lähteiden ominaispiirre on, että niitä on runsaasti ja ne ovat pirstaloituneita. Tieto jakautuu useille eri verkkosivuille, ja tarvittavan tiedon löytäminen vaatii lukuisten linkkien klikkailua. Olennaisen tiedon suodattaminen opinnäytetyön näkökulmasta oli pidettävä mielessä koko ajan. Rajasin lähdeaineistoa myös ajan mukaan ja käytin aina uusimpia versioita lähteistäni. En ole merkinnyt Microsoftin lähteiden yhteyteen niiden päiväyksiä, vaan erottelen lähteet vuosiluvun ja kirjaimen avulla. Microsoftin lähteet päivittyvät jatkuvasti ja on oletettavaa, että jossakin vaiheessa säilytyksen palvelutasojen hinnoitteluun ja periaatteisiin tulee muutoksia.

Käytän lähteinä myös Microsoft Azuren sertifiointeihin valmistavia teoksia liittyen kursseihin *AZ-900 Microsoft Azure Fundamentals* ja *Implementing Microsoft Azure Architect Technologies: AZ-303*. Näistä löytyy kootusti tietoperustaa lokien analysoinnista ja Azuren säilytyskustannuksista. Erityisesti Jim Cheshiren äänikirja *AZ-900* -kurssin harjoittelumateriaaliksi on suositeltava perusteos, joka lähtee liikkeelle aivan perusasioista.

Yaser Mansouri ja Abdelkarim Erradi esittelevät vuonna 2018 julkaistussa artikkelissaan “*Cost Optimization Algorithms for Hot and Cool Tiers Cloud Storage Services*” laskenta-algoritmin, johon arvot syöttämällä saisi selville tilanteet, joissa viileällä palvelutasolla säilyttäminen olisi kustannustehokasta verrattuna kuumalla palvelutasolla säilyttämiseen. Artikkelin jatko-osa ilmestyi *Journal of Systems and Software* -julkaisussa vuonna 2020. Vaikka käytän artikkeleita lähteinäni, pohjaan opinnäytetyöni laskelmat niissä esitettyjen algoritmien sijasta (kyseessä ovat matemaattiset laskukaavat) Microsoft Azure -verkkosivuston kustannuslaskuriin.

Luku- ja kirjoitus -operaatioihin perustuvan algoritmin avulla voidaan ohjelmoida kullekin tieto-objektille sopivin säilyttämiseen liittyvä palvelutaso ja tarvittaessa tason muutos, jos objektiin kohdistuvan käytön määrä muuttuu (Mansouri,& Erradi 2018, 622). Mansouri ja Erradi ovat suunnitelleet myös koneoppimismallia oikean palvelutason valintaan. Lisäksi he kertovat artikkelissaan ajatuksesta, jonka mukaan dataan voisi liittää tunnisteita ja niitä voisi luokitella riippuen siitä, minkälainen ennuste niiden tulevaan käyttöön pätee. (Mansouri & Erradi 2020, 13).

Lukuisilla IT-alan yrityksillä on verkkosivuillaan tietoa tietoaltaiden säilytyskustannusten hallinnasta, mutta näitä lähteitä käytän opinnäytetyössäni harkiten niiden markkinointinäkökulman vuoksi. Kaupallisista toimijoista Ari Hovi Oy:n ja TietoEvryn blogit toimivat hyvin käsitteiden ja termien avaamisessa.

Aikaisempia AMK-opinnäytteitä tai pro graduja opinnäytetyöni aiheesta on hyvin rajallinen määrä, eivätkä ne suoraan käsittele opinnäytetyöni asetelmaa. Saku Junni on vuonna 2020 Laurea Ammattikorkeakoululle laatimassaan opinnäytetyössä avannut tietovarastoinnin hyviä käytänteitä ja siihen liittyen kustannusten hallintaa.

Tietoperusta tarjoaa pohjan empiirisessä osassa tekemilleni kustannuslaskelmille ja sen avulla pystyn esittämään erilaisia hyödyllisiä näkökohtia pohdittaessa sopivinta tiedon säilyttämisen palvelutasoa.

2.2 Azure Data Lake Storage Gen 2: tiedon säilytyksen hinnoittelun periaatteet

Kustannuksissa säästäminen on yksi keskeinen syy sille, miksi tietoa viedään pilviympäristöön. Käyttöön ja säilytykseen pohjautuvien hinnoittelumallien kanssa on kuitenkin tasapainoteltava sen kanssa, milloin tiedon säilyttämiseen liittyvän palvelutason muuttaminen on taloudellisesti hyödyllistä. Kustannusoptimoinnin nimissä tietoa kannattaa säilyttää sen palvelutason mukaisesti, mille se kuuluu. (Mansouri, Y., Erradi, A. 2020, 13).

Oikean palvelutason, tai ylipäätään tiedon säilytysratkaisun, valintaan vaikuttavat datan käyttötarkoitus, uuden datan sisään tuonnin tiheys lähdejärjestelmästä, tiedon formaatti ja koko (Reis & Housley 2022, 41; 228). Kylmälle palvelutasolle kuuluu data, jota säilytetään *compliance* -tarkoituksiin – esimerkiksi lainsäädännön vaatimuksesta – tai varmistuksina. Dataan ei ole tarvetta kohdistua säännöllisiä aktiviteetteja. (Reis & Housley 2022, 41; 228).

Kuuma palvelutaso, *hot tier*, on optimaalisin usein käytettävälle datalle. Tyypillisesti kuumalla palvelutasolla säilytettävään dataan kohdistuu runsaasti käyttöä, eli useita kertoja päivässä ja tietyissä tapauksissa jopa useita kertoja sekunnissa. Tieto pitää tällöin olla nopeasti saavutettavissa. (Reis & Housley 2022, 41; 228). Kuuma palvelutaso soveltuu myös ensimmäisen vaiheen tallennukseen ns. *staging* -datalle (tilapäiseen säilytykseen), joka tuodaan sisään tietoaltaaseen, mutta se jatkaa matkaansa tietyn lyhyehkön ajan jälkeen viileälle tasolle. (Microsoft 2023d).

Viileä palvelutaso, *cool tier*, soveltuu parhaiten satunnaisesti käytettävän datan säilytykseen. Dataa tulee säilyttää vähintään 30 päivän ajan. Tyypillisiä viileän palvelutason säilytykseen soveltuvia ”tapauksia” ovat varmistukset, tarvittaessa käyttöön otettava data tai suuret tietomäärät, jotka vaativat

kustannustehokkaan säilytysratkaisun, mutta eivät enää ole ihan uusinta tietoaltaaseen sisään tulevaa dataa. (Microsoft 2023d).

Kylmä palvelutaso, *cold tier*, sopii harvemmin tarvittavalle datalle, joka kuitenkin pitää tarvittaessa saada käyttöön kohtuullisessa ajassa. Säilytysaika kylmällä palvelutasolla tulee olla vähintään 90 päivää. Toisin kuin edellä mainitut palvelutasot, arkistointiin tarkoitettu *archive tier* on offline-tilainen. Data on noudettavissa arkistointiin soveltuvalta palvelutasolta tuntien kuluessa. Tällainen data tulee tallettaa vähintään 180 päiväksi ja sitä ei pitäisi juurikaan aktiivisesti käyttää.

Tietojen säilyttämiselle on kuumaa palvelutasoa lukuun ottamatta määritelty erilainen minimiaika, joka on tasosta riippuen 30, 90 tai 180 vuorokautta. Jos tiedot mennään poistamaan kyseiseltä palvelutasolta ennen tuon ajan täyttymistä, siitä seuraa rangaistusmaksu. (Microsoft 2023d).

Perusperiaate Azuren hinnoittelussa on, että viileämpi palvelutaso on säilytyksen kannalta edullisempaa, mutta jos dataan kohdistuu luku- tai muita operaatioita, niiden hinta on viileällä palvelutasolla kalliimpi kuin kuumalla palvelutasolla säilytettävien tietojen kohdalla. (Microsoft 2023d).

Azure-palveluiden hintaa ennakoidaan kustannuslaskurilla (*Azure pricing calculator*, ks. Liite 1), jossa säilyttämisen kustannuksia arvioidaan tallennustilin tasolla (*Storage account*). Esimerkeissäni käytän hintaan vaikuttavina muuttujina säilytyskapasiteetin kokoa (gibitavut, tebitavut) ja luku-operaatioiden määrää. Lisäksi joudun huomioimaan kirjoitus-operaatioiden määrän, sillä Microsoftin mukaan palvelutason tyypin vaihto lasketaan kirjoitus-operaatioiksi. (Microsoft 2024a). Hinnat valitsen näytettäviksi euroina.

Tosiasiassa ADLS Gen2 -hinnoitteluun vaikuttavat tarvittavan säilytyskapasiteetin ja operaatioiden määrän vuoksi myös alueen valinta (*region*), tallennustilin ja säilytysratkaisun tyyppi, datan siirtomaksut, maksutavan valinta (*pay-as-you-go* tai varattu kapasiteetti) sekä datan varmistaminen. (Slingerland 2023).

Azure Data Lake Storage Gen 2 -hinnoittelun perustaksi kustannuslaskuriin syötetään seuraavat tiedot:

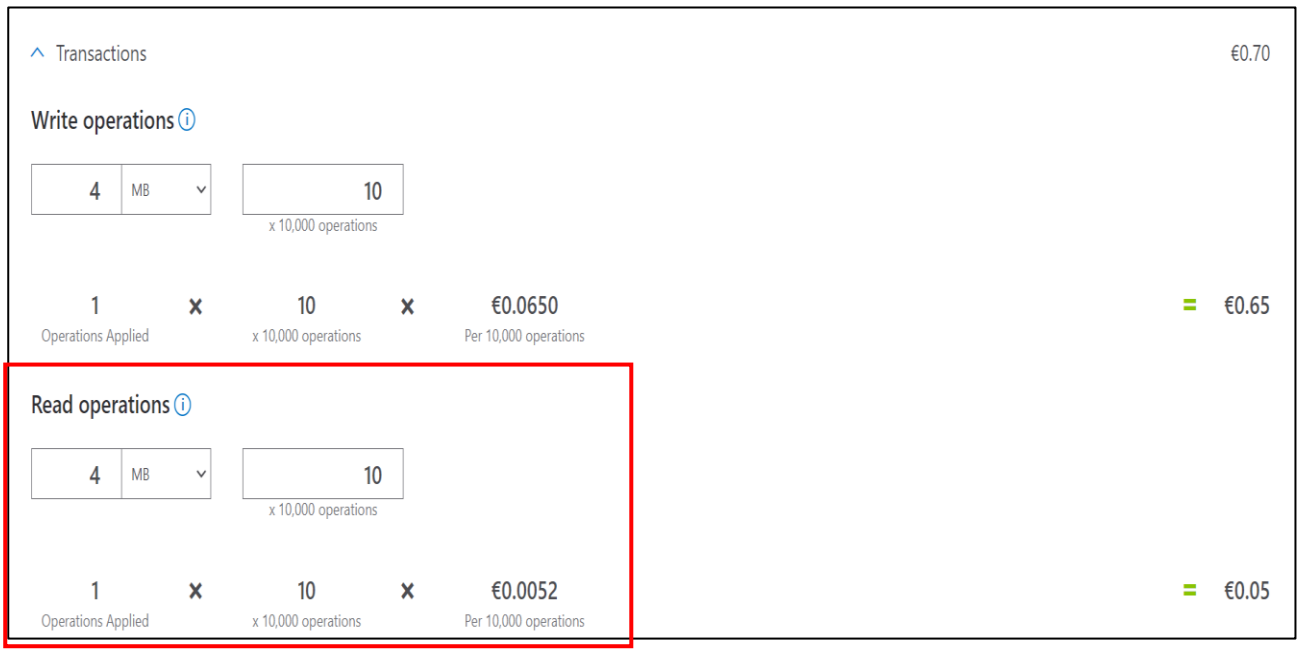
- Alue (*region*): tarjolla on lukuisia eri puolilla maailmaa sijaitsevia palvelimia. Käytän kaikissa käyttötapauksissani vertailtavuuden vuoksi alueena Länsi-Eurooppaa (*West Europe*).
- Tyyppi (*type*): Tähän valitaan Data Lake Storage Gen 2.
- Tilauksen taso (*tier*): *Premium* tai *Standard* -tilaus. Tämä on eri asia kuin säilytyksen palvelutaso eli *access tier.* Käyttötapauksissani käytän tässä arvoa *Standard*.
- Tallennustilin tyyppi: (*storage account type*): vaihtoehtona on vain *General Purpose V2.* Tämä tallennustilin tyyppi soveltuu useimpiin käyttötarkoituksiin, ja nimen omaan blob storage -tyyppisille tiedon säilytysratkaisuille (Cheshire 2022, luku 2.3).

- Palvelutason valinta: tässä määritellään kuuma, viileä, kylmä tai arkistointiin tarkoitettu säilytyksen palvelutaso.
- Säilytettävien tietojen varmennus: tieto siitä, varmistetaanko tietojen säilyminen paikallisesti (*Locally redundant storage, LRS*), vyöhykkeellisesti (*Zone-redundant storage, ZRS*), maantieteelliseen alueeseen perustuvasti (*Geo-redundant storage, GRS* tai *Geo-zone-redundant storage, GZRS*) tai maantieteelliseen alueeseen perustuvasti, mutta vain-luku -varmennuksena (*Read access-only redundant storage, RA-GRS tai RA-GZRS*) (Microsoft 2024b). Käytän opinnäytetyöni laskelmissa arvoa LRS.
- Tiedostorakenne: joko hierarkkinen tai litteä nimiavaruus (*hierachical* tai *flat namespace*). ADLS Gen 2 tarjoaa kahdenlaista mahdollisuutta tietojen järjestämiseen. Hierarkkisessa nimiavaruudessa hakemistot ja tiedostot on nimetty uniikisti ja hakemiston tai tiedoston uudelleen nimeäminen tehdään metadatan avulla. Tässä vaihtoehdossa data on tiedostomaisessa rakenteessa, kansioina ja tiedostoina. Litteä nimiavaruus on rakenteeton lista talletetuista kohteista. Azuren tallennustilillä litteä nimiavaruus on oletusarvoinen. Hierakkinen nimiavaruus tuottaa kustannuksia metadatan säilyttämisestä, jos sellaista haluaa lisätä blobeihin. Metatiedon kustannus lasketaan jokaiselle säilytettävälle tiedostolle - tai blobille - siten, että tiedoston kokoon lisätään 512 tavua + tiedoston nimen koko + tiedoston ominaisuuksien (*properties*) koko. (Azure Storage 2023). Käytän opinnäytetyöni käyttötapauksissa hierarkkista nimiavaruutta.
- Tallennuskapasiteetin koko. Tämän arvion saan selville Azure Storage Explorerista, jossa tietoaltaan tiedot näkyvät kansioituina.
- Tieto siitä, valitaanko *pay as you go* -hinnoittelumalli, vai käytetäänkö yhden tai kolmen vuoden varattua kapasiteettia. Varatussa mallissa kustannuksia olisi hyvä pystyä ennakoimaan, jotta tarpeellinen kapasiteetti tulisi varatuksi kerralla oikein (Microsoft 2024b). Käyttötapauksissani käytän *Pay as you go* -valintaa ja syötän tiedot kuukausitasolla.

Pay-as-you-go -hinnoittelumalli soveltuu tilanteisiin, joissa datan käsittelyyn tarvitaan joustavuutta ja skaalautuvuutta, erityisesti jos käyttöä ei voida täysin ennakoida (Guruswamy 2024). Varattu säilytyskapasiteetti maksetaan etukäteen joko yhdeksi tai kolmeksi vuodeksi. Näin voidaan säästää jopa 34% kuuman tai viileän palvelutason kustannuksissa. Arkistointiin tarkoitetulla palvelutasolla luvattu säästö on 17%. Varatun kapasiteetin kohdalla on syytä käyttää harkintaa, sillä vuodeksi tai kolmeksi sitoutuminen tiedon säilyttämiseen tietyllä palvelutasolla on riski, eikä säästö lopulta ole erityisen suuri verrattuna *pay-as-you-go* -hintaan. Lisäksi varattu kapasiteetti vaatii kullakin palvelutasolla vähintään 90,9 tebitavun/100 teratavun datan. Dataan kohdistuvat operaatiot laskutetaan aina *pay-as-you-go* -mallilla. (Veritas Technologies 28.9.2022).

Seuraavaksi Microsoftin kustannuslaskurissa siirrytään arvioimaan kirjoitus- ja luku-operaatioita (*write operations, read operations*), datan noutoa (*retrieval*) ja muita mahdollisia operaatioita. Operaatioiden laskutus perustuu 3,81 mebitavun/4 megatavun sääntöön. Azure ”pilkkoo” kirjoitus- ja luku-operaatiot 4 megatavun kokonaisuuksiksi. Vaikka operaatio olisi pienempi kuin tämä tavumäärä, laskutus on silti 4 megatavuun perustuva. (Azure Storage 2023). Microsoft ei kerro, onko operaation koko todellisuudessa 3,81 mebitavun suuruinen, mutta johdonmukaisuuden vuoksi ilmaisen tämänkin luvun myös binäärisessä muodossa. Kerron luvussa 3.1 tarkemmin näistä tiedostokokojen määritelmistä.

Operaatioiden odotettavissa oleva lukumäärä kuukaudessa sijoitetaan kustannuslaskurissa ”oletuskoon” oikealla puolella olevaan kenttään, jossa alla olevassa kuvassa 2 on luku 10.



Kuva 2. Azuren kustannuslaskurin kirjoitus- ja luku -operaatioiden osuus (Microsoft 2024b). Korostettuna luku-operaatiot. Mebitavuina yhden operaation koko on 3,81.

Kustannuslaskurissa kysytään lisäksi datan kyselyjä nopeuttavan *Query Acceleration* -palvelun ja siihen liittyvän datan käytön arvioita *(Data Returned, Data Scanned*). Seuraavana vuorossa ovat iteratiiviset luku- ja kirjoitusoperaatiot. Tämä tarkoittaa tiettyjä rajapinnan kautta tehtäviä *List* – ja *Rename* -operaatioita, joissa joudutaan käymään läpi kaikki alikansiot tai tiedostot kansiossa. Kustannuslaskuriin syötetään myös kaikki muut tiedossa olevat operaatiot, pois lukien poistaminen, *delete*, joka on ilmaista. Näihin sisältyvät datan nouto (*retrieval*), kirjoittaminen (*data write*, joka on eri asia kuin kirjoitusoperaatio) ja metatiedon säilyttämisen tarve. (Microsoft 2024b). Arkistointiin tarkoitetun palvelutason kohdalla yksi iteratiivisen luku-kustannuksen tyyppi on nopeampi luku-

operaatio (*archive high priority read*) ja yksi tiedonnoudon kustannustyyppi niin ikään nopeampi *archive high priority retrieval*. (Microsoft 2023j)

Jos tiedostoihin lisätään metadataa, jota hierarkkisessa tiedostorakenteessa vaaditaankin, sen säilytyskustannukseksi tulee arvioida 512 tavua + esimerkin omainen tiedoston nimen koko + esimerkin omainen tiedoston ominaisuuksien koko ja kertoa se tiedostojen määrällä (Azure Storage 2023). En ole opinnäytetyössäni laskenut kustannuksia metatiedolle. Näiden kustannusten arviointiin tarvittavat tiedot kyllä saisi selville Azure Storage Explorerista tiedostoja ja niiden ominaisuuksia tarkastelemalla.

Laskurin lopussa tiedustellaan vielä asiakastuen tason tarvetta ja lisenssisopimusta. Myös kehitys- ja testiympäristöjen kulut voidaan arvioida mahdollistamalla tämä vaihtoehto. Kustannuslaskelman voi viedä Exceliin sivun alareunassa olevasta *Export*-painikkeesta. (Microsoft 2024b).

Arkistointiin tarkoitettu säilytyksen palvelutaso ja kylmä palvelutaso voidaan lähdekirjallisuuden mukaan määrittää vain ”blobin” tasolla (Güntert 22.1.2023). Azuren kustannuslaskuri kuitenkin mahdollistaa ainoastaan tallennustilin tasoiset laskelmat ja se antaa kustannusarvion kaikille säilytyksen palvelutasoille. Arkistointiin tarkoitetun palvelutason oletushinta on sama, tehdään kustannusarvio sitten ADLS Gen2- tai sille rinnakkaiselle ”blobeja säilyttävälle” *Block Blob Storage* -palvelulle. Oletuksena palvelutaso periytyy tallennustilin tasolta. Samalla tallennustilillä voi olla useamman palvelutason dataa. (Microsoft 2023a).

3 Välineet kustannusten selvittämiseen ja hallintaan

Esittelen tässä luvussa työkalut, joita tarvitsen hakiessani tietoa Azure Data Lake Storage Gen 2 -tietoaltaan käytöstä Microsoftin kustannuslaskuria varten. Lisäksi kerron tietojen siirtämisestä palvelutasolta toiselle joko elinkaarenhallinnan sääntöjen avulla tai erilaisilla teknisillä vaihtoehdoilla.

Azure Storage Explorer on ”ikkuna” tietoaltaaseen tallennettuun dataan. Sen avulla voidaan etsiä tiedostokansioiden tai yksittäisten tiedostojen kokoja. Etsin Azure Storage Explorerin avulla luvussa 4 esittelemiäni käyttötapauksia varten tiedostokansioiden muodostamia kokonaisuuksia, joiden oletan kuuluvan samalle palvelutasolle.

Lokitiedot ovat avainasemassa tietoaltaan käytön ymmärtämisessä. Kerron Microsoftin Log Analytics -palvelusta ja esittelen sen KQL-kielisen hakulauseen, jolla saadaan selville kustannuslaskuria varten tarvittavat luku-operaatioiden määrät ja niiden keskimääräiset koot. Konkretisoin lokien kautta selville saamieni lukujen käyttöä osana kustannuslaskentaa niin ikään luvun 4 käyttötapauksissa.

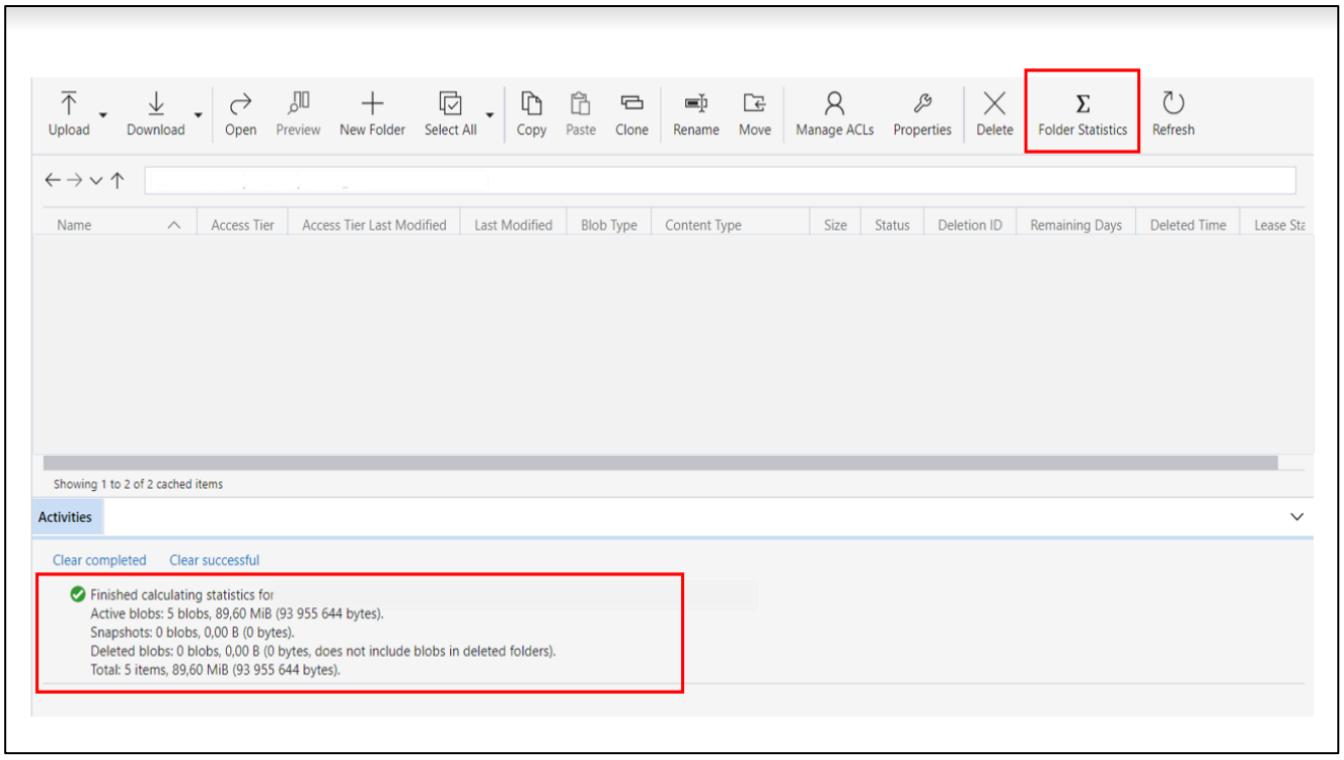
3.1 Azure Storage Explorer näkymänä tietoon ja säilytettävien kohteiden koon etsinnässä

Azure Storage Explorer on verkosta ladattava työpöytäsovellus, joka mahdollistaa Azuren pilvipalveluihin tallennetun tiedon tarkastelemisen. Käyttö edellyttää Azure-tilausta (*subscription*), johon Storage Explorerista luodaan yhteys. Azure Storage Explorer toimii Linux-, iOS-, ja Windows -käyttöjärjestelmillä. Sen avulla voidaan luoda uusia säiliöitä (*containers*), tuoda tiedostoja, kopioida tietoja säiliöstä toiseen, myöntää Shared Access Signature (SAS) -käyttövaltuuksia ja sillä voi muuttaa palvelutasoa. "*Change access tier*" -toiminto voidaan tehdä klikkaamalla hiiren oikealla painikkeella vaikka yksittäisen blobin tasolla (Cheshire 2022, luku 2.3).

Azure Storage Explorer näyttää ADLS Gen 2:een tallennetut kohteet kuin tiedostohakemistona kansioineen. Näiden kansioiden sisällä voi olla toisia kansioita ja lopulta itse tiedostot. (Microsoft 2024c).

Tiedostokansioihin voidaan viitata skeemoina, ja ne voivat yhdessä muodostaa tietokannan. Skeema -käsitettä käytetään tietosisällön, tietolähteen tai käyttöoikeuksien perusteella ryhmitellystä tiedosta, joka voi Azure Storage Explorerin näkymässä kuulua samaan tiedostokansioon. Olennaista säilytyksen palvelutason valinnan kannalta on selvittää, mitkä tiedot muodostavat sellaisia kokonaisuuksia, että niitä kannattaa säilyttää samalla palvelutasolla.

Tiedostojen koot haetaan Azure Storage Explorerilla esiin siten, että navigoidaan haluttuun kohteeseen, valitaan se aktiiviseksi ja klikataan valintapaneelin kohtaa *Folder Statistics*. Tiedostojen koot tulevat näkyviin käyttöliittymän alareunaan. Tarkastelen opinnäytetyössäni aktiivisia ”blobeja” eli tiedostokokoja kohdassa *Active blobs*. Vaihtoehto *Snapshots* tarkoittaa vain-luku -kopioita, joita voidaan ottaa mistä tahansa blobista. (Microsoft 2023k).

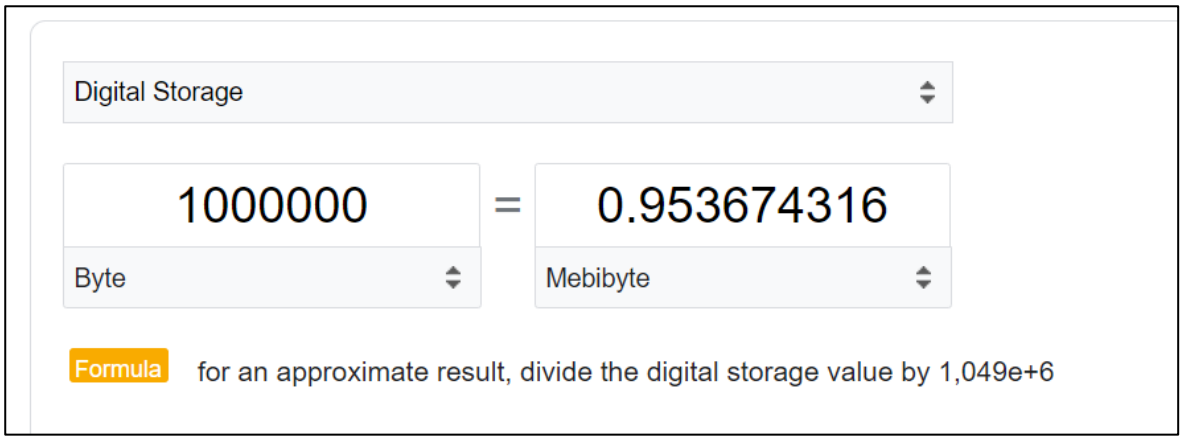


Kuva 3. Tiedostokokojen tarkasteleminen Azure Storage Explorerin avulla.

Azure Storage Explorer ilmoittaa säilytettävien tietojen koot mebi-, gibi- tai tebitavuina. Microsoft Azuren kustannuslaskurissa tarvittavan tallennuskapasiteetin koko on annettava mega-, giga- tai teratavuina. Syötän tiedot laskuriin johdonmukaisesti binäärisinä yksiköinä eli bi-loppuisina yksiköinä ja muunnan Microsoftin käyttämät yksiköt sellaisiksi. Yksiköiden muunnoksiin on internetissä tarjolla laskureita, esimerkiksi sivustolla gbmb.org (Gbmb 2024). Myös Googlelta (www.google.fi) löytyy laskuri, jolla tehdä muunnokset. Laskuri avautuu, kun syöttää Googlen hakukenttään vaikkapa sanat ”Bytes to Mebibytes”.

Kibitavut, mebitavut, gibitavut ja tebitavut kuvaavat tiedon säilytystä siinä missä monelle tutummat kilo-, mega-, giga- ja teratavutkin. Näiden käsitteellinen ero selittyy sillä, että ensin mainitut ovat binäärisiä – kuten tietokoneiden datakin – ja jälkimmäiset ovat metrisiä yksiköitä. Binääristen yksiköiden käytöstä linjattiin vuonna 1998 IEC:n (*International Electrotechnical Commission*) päätöksellä. Kilotavu metrisenä yksikkönä vastaa 1000 tavua, kun taas kibitavu on 1024 tavua. Binääriset

yksiköt ovat yleistyneet kun datan määrä ja sen säilytyskapasiteetti yleisesti on kasvanut, ja tämän myötä erilaisten mittayksiköiden välinen ero korostuu. (Donnelly 2021).



Kuva 4. Googlen laskuri tiedon säilytysyksiköiden muuntamiseen. Laskurin käytössä on huomioitava, että erotinmerkkinä käytetään pistettä pilkun sijaan. Jos muunnettavana on esimerkiksi 3,5 megatavua, tämä merkitään muodossa 3.5.

3.2 Microsoft Log Analytics luku-operaatioiden analysoinnissa

Lokien avulla on mahdollista tunnistaa datasta odottamattomia tapahtumia, suhteita, syitä ja seurauksia tai trendejä ja kaavamaisuuksia. Vanhaa lokitietoa voidaan käyttää tulevaisuuden suuntaviivojen hahmottamiseen. Lokitiedon tarkastelusta on erityistä hyötyä kyberturvallisuudelle ja monimutkaisen IT-infrastruktuurin hallinnalle, suorituskyvyn optimoinnille ja virheiden selvittelylle. (Gillespie & Givre 2021, luku 2). Lokien tarkastelu on osa pilviympäristön monitorointia. Myös kustannusten hallinnan kannalta on tärkeää tietää, mitä ympäristössä tapahtuu. (Hargreaves & Zaal 2020, luku 1).

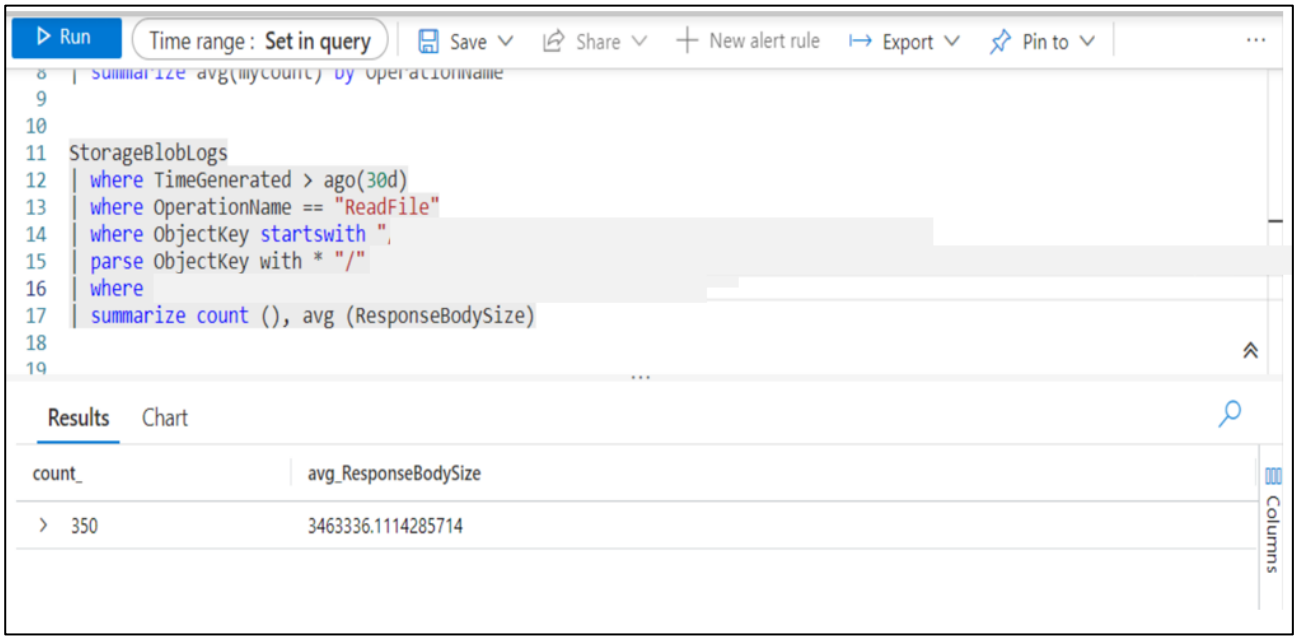
Log Analytics on Microsoftin asiakasportaalissaan (portalazure.com) tarjoama palvelu, jonka avulla voidaan kysellä ja analysoida Azuren keräämiä lokeja. Log Analytics löytyy asiakasportaaliin sisään kirjautumisen jälkeen Monitor-valikosta. (Microsoft 2023f). Opinnäytetyössäni käytän Azuren lokeista löytyviä luku-operaatioiden määriä syöttääkseni ne kustannuslaskuriin ja arvioidakseni erilaisten säilytyksen palvelutasojen optimaalisuutta kussakin käyttötapauksessa. Takautuvien operaatioiden määrää voidaan käyttää arvioidessa tulevaisuuden luku-operaatioita, mikäli toiminnan oletetaan jatkuvan samankaltaisena.

Azure Data Lake Storage Gen 2 -lokeja kysellään StorageBlobLogs -määritteellä. Ennen kyselyä on varmistettava, että se kohdistuu oikeisiin tallennustileihin (*Storage accounts*)..

Selvittääkseni luku-operaatioiden määriä kustannusarviointia varten kuukauden ajalta, tarvitsen tiedon a) luku-operaatioiden määrästä viimeisimmältä 30 päivältä ja b) keskimääräisen luku-operaation koon viimeisimmältä 30 päivältä. Koko ilmoitetaan tavuina.

Oheinen KQL-kysely tarjoaa tarvitsemani vastaukset:

```
StorageBlobLogs
| where TimeGenerated > ago(30d)
| where OperationName == "ReadFile"
| where "[määritellään tiedostokansiota tarkemmin]"
| summarize count (), avg (ResponseBodySize)
```



Kuva 5. Luku-operaatioiden määrän ja keskimääräisen koon hakeminen Log Analyticsista. Tarkasteluun haluttavaa kohdetta rajataan esimerkiksi ObjectKey- tai Container -määritteillä.

Hyödyllistä voi olla myös hakea tiettyjen applikaatioiden toimintaan liittyviä operaatioita. Hakuehtona tämä on `RequesterAppID` (Microsoft 2024d). Applikaatio voi olla esimerkiksi Azure Synapse Analytics, jonka tuottamilla operaatioilla on oma ID-numeronsa.

Luku-operaatioita summatessa ja määriä analysoidessa kannattaa tarkastella, mistä kaikesta kyseisiä operaatioita muodostuu eli miksi applikaatio tekee kutsuja säilytyspalvelun suuntaan. Tietoaltaan dataan voi kohdistua erilaisia ajoja ja skannauksia, joiden tarkoitus on hyvä kustannussäästöjen nimissä selvittää. Tietoaltaaseen kohdistuvat operaatiot kannattaa käydä läpi applikaatio kerrallaan ja selvittää, miten ne käyttävät resursseja (Puntanen & Virkkunen 5.3.2024).

3.3 Azure Lifecycle Management -työkalu ja muut vaihtoehdot säilytyksen palvelutason muuttamiseen

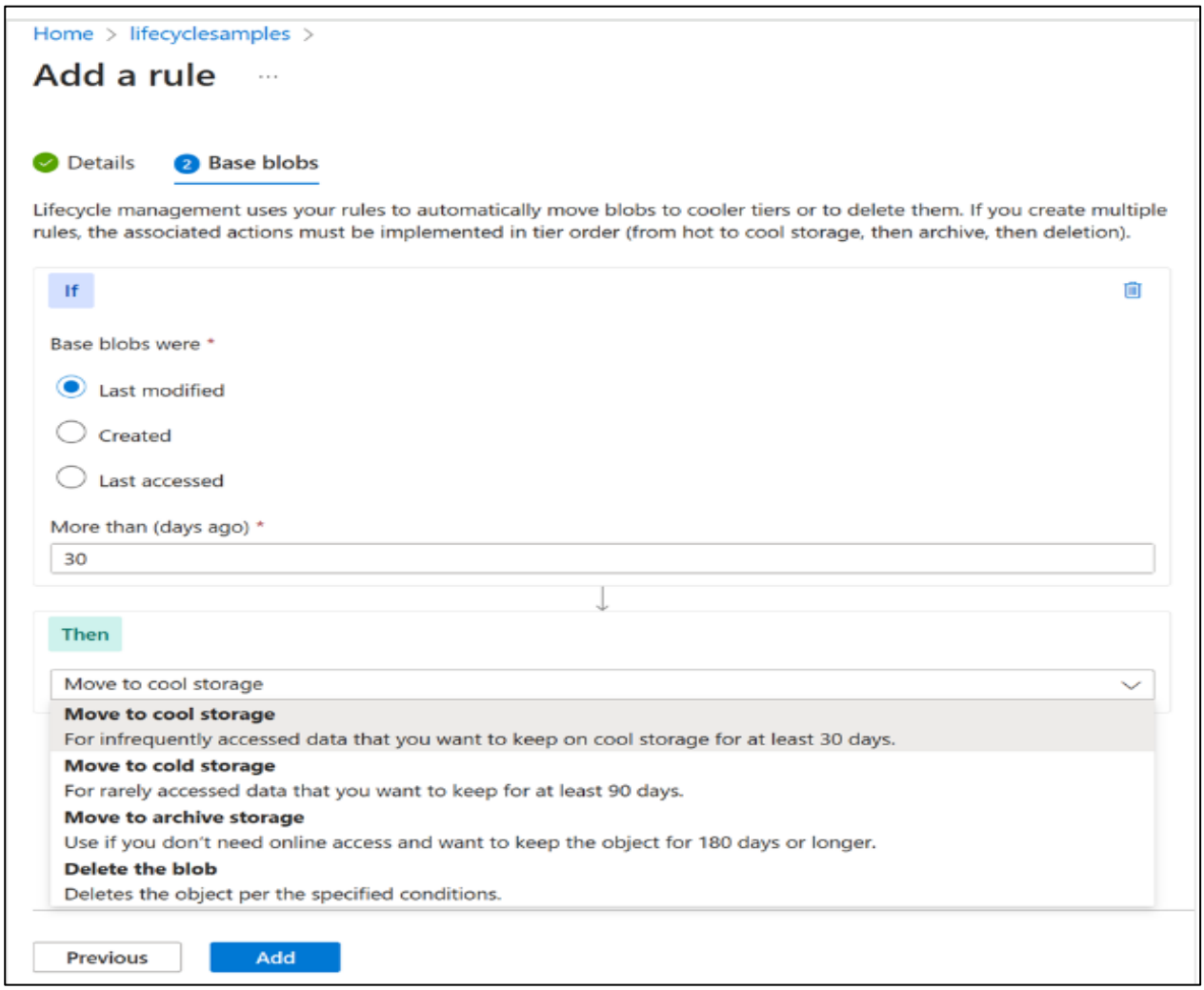
Automaattisesti toteutetut tiedon elinkaaren hallinnan säännöt helpottavat kustannusten hallintaa. Koska säilytyksen palvelutason mukainen hinnoittelu pohjautuu tiedon käyttötiheyteen, ja uusin data on useimmiten käytetympää kuin vanha, on suositeltavaa hyödyntää säilytyspalveluihin kytkettyjä säännöstöjä (*policies*), joiden mukaan esimerkiksi 180 vuorokautta vanhempi, koskematon data, siirtyy viileämmän palvelutason piiriin. (Reis & Housley 2022, 229).

Azuren Lifecycle Management -työkalu mahdollistaa säilytyksen palvelutason muutoksen tai datan poistamisen perustuen tietojen viimeisimpään käyttöön. Säännöt voidaan ulottaa kokonaiseen tallennustiliin, valittuihin säiliöihin tai blobeihin. Käytännössä elinkaaren hallinnan säännöstöt luodaan ja ylläpidetään joko Azuren asiakasportaalissa, PowerShell- tai AzureCLI -komennoilla tai Azure Resource Manager -mallipohjilla (*ARM templates*). Palvelu itsessään ja elinkaarenhallinnan säännöstöjen luominen, samoin kuin tietojen poistaminenkin (*delete*) on maksutonta. Kustannuksia tulee kuitenkin *SetBlobTier* -tyyppisistä rajapintakutsuista (*API calls*) tai muusta operaatiosta, jolla palvelutasoa muutetaan. Myös muut Azuren palvelut, kuten tietoturvaan liittyvä *Microsoft Defender for Storage* saattaa muodostaa kustannuksia säilytyksen palvelutason muutoksista. (Microsoft 2023c).

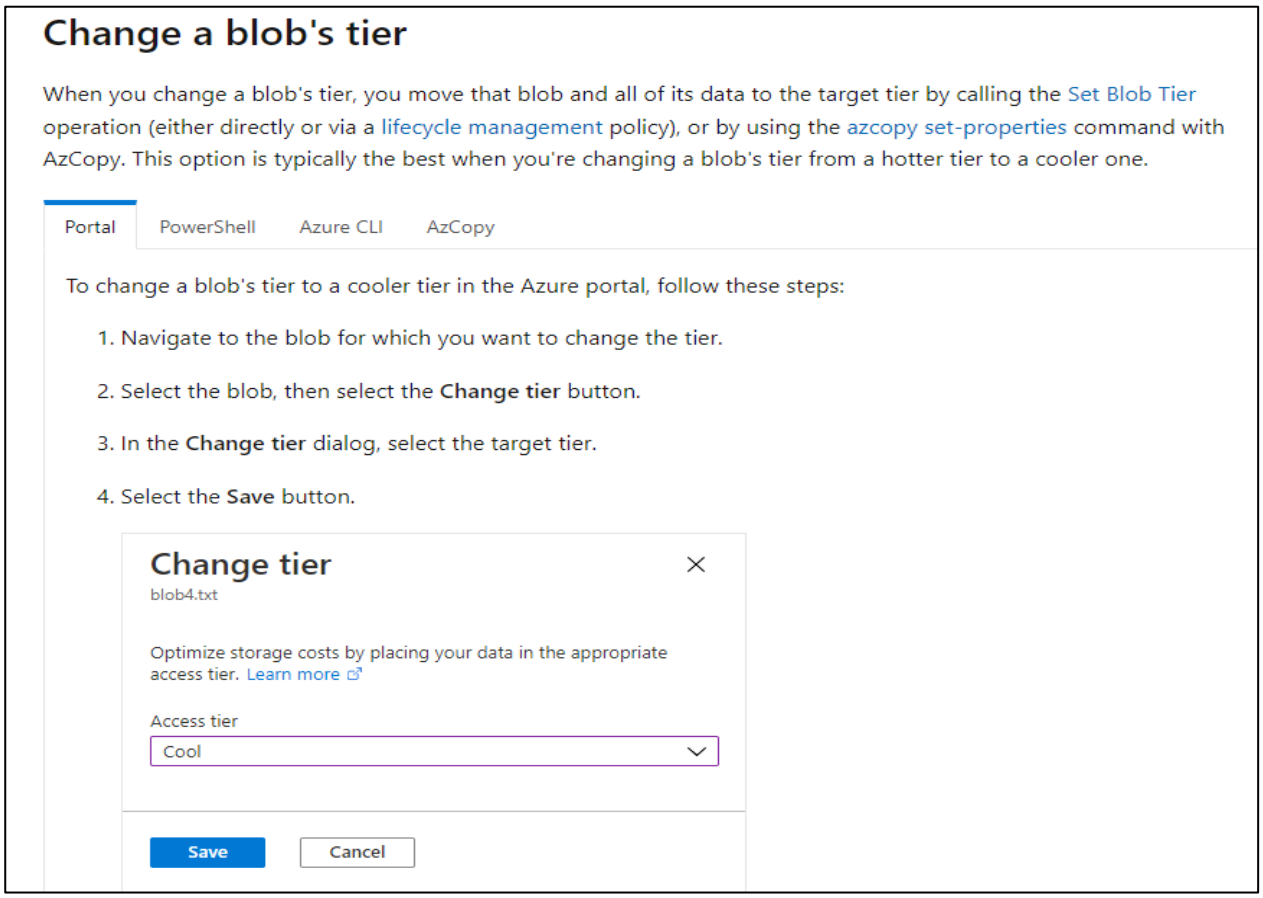
Aikaleimana, jonka perusteella säännöt muuttavat tiedon palvelutasoa tai johtavat tiedon poistamiseen, voidaan käyttää tiedon viimeisintä muokkausajankohtaa tai viimeisintä käyttöajankohtaa (*last access time*). Viimeisimmän käyttöajankohta -tiedon kerääminen blobeille täytyy erikseen sallia. Näin voidaan esimerkiksi luoda sääntö, että kuumalla palvelutasolla säilytettävät blobit siirtyvät viileälle palvelutasolle, kun niihin ei ole kohdistunut 30 päivään yhtään luku- tai kirjoitusoperaatiota. (Microsoft 2023c).

Jos Azuren Lifecycle Management -työkalun tarjoamaa, automatisoitavaa säilytyksen palvelutason muutosta ei haluta jostakin syystä käyttää, palvelutason muuttaminen tapahtuu joko Azure Storage Explorerissa, Azuren asiakasportaalissa, PowerShell -, Azure CLI- tai AzCopy -komennoilla. Kun dataa halutaan siirtää massoittain tiedostokansioissa vaikkapa arkistointiin tarkoitetulle palvelutasolle, tämä siirtäminen on suositeltavaa tehdä eräajona hyödyntäen esimerkiksi *AzBulkSetBlobTier* -komentoa. (Microsoft 2023l).

AzCopy on komentorivityökalu, jonka avulla voidaan kopioida blobieja ja tiedostoja Azuren säilytyspalveluissa ja niiden välillä. Yhdellä operaatiolla voidaan myös kopioida kokonaisia hakemistoja. AzCopylla voidaan niin ikään kirjoittaa automatisoitavia skriptejä. (Cheshire 2022, luku 2.3).



Kuva 6. Säilytyksen palvelutasojen muutosten automaattiseen hallinnointiin tarkoitettu Microsoftin Lifecycle Management -työkalu Azuren asiakasportaalissa (Microsoft 2023m)



Kuva 7. Näkymä siihen, miltä säilytyksen palvelutason muuttaminen näyttää Azuren asiakasportaalissa. Myös PowerShell, Azure CLI ja AzCopy näkyvät vaihtoehtoina. (Microsoft 2023l)

4 Palvelutason optimointia esimerkkien avulla

Käyttötapausteni avulla testaan, minkälaisia tuloksia Microsoftin hinnoittelu eri palvelutasoille erilaisilla kustannusmuuttujilla tuottaa. Esitietoina olen opinnäytetyöni luvussa 3 esittelemilläni välineillä (Azure Storage Explorer ja Microsoftin Log Analytics) etsinyt käyttötapauksiani varten tiedot tarvittavasta säilytyskapasiteetin koosta, luku-operaatioiden määrästä ja kyseisten operaatioiden keskimääräisestä koosta.

Käyttötapausten esimerkit liittyvät analytiikkaan ja niistä varsinkin viimeisessä käsitellään huomattavia datamääriä. Todellisuudessa yhden tallennustilin alla voi olla useilla eri palvelutasoilla säilytettävää dataa, mutta esimerkeissäni tallennuskapasiteetti on samalla koko tallennustilin tarvitsema kapasiteetti.

Luvussa 2 esittelemässäni Mansourin ja Erradin algoritmissa tietojen “elinaika” säilytyksessä vaihtelee yhden ja kolmen kuukauden välillä. Tuolla aikavälillä useimpien tieto-objektien luonne muuttuu siten, että ne on kustannustehokkaampaa siirtää viileämmälle palvelutasolle. (Mansouri & Erradi 2018, 628). Käytännössä säilytysajat voivat kuitenkin olla lainsäädännöstä tai muusta toimintaympäristöstä johtuen huomattavasti pidempiä ja käyttö aktiivisena jatkua kauemmin kuin kuukausia. Tämä vaikuttaa luonnollisesti palvelutason valintaan.

4.1 Käyttötapaus 1: Harvoin käytettävät, arkistoitavat tiedot

Tällaisen käyttötapauksen kaltaiset tiedot voivat olla säilytyksessä tietoaltaassa esimerkiksi lainsäädännön vaatimusten tai sopimuksen ehtojen täyttämiseksi. Tiedoilla voi olla hyvin pitkä säilytysaika, mutta niitä tarvitaan harvoin. Käyttötapauksessa kyseessä ovat kertaladatut, jopa kymmenien vuosien ajan säilytettävät tietokantataulut. Tietoja säilytetään pohjadatana raporteille, joille dataa ladataan silloin tällöin.

Näiden arkistointitarkoitukseen säilytettävien tiedostojen koko saadaan selville Azure Storage Explorerin kautta. Dataa on tiedostokansioissa ylöspäin pyöristettynä 219,15 tebitavun/240 teratavun verran. Blobeja on 180 kappaletta. Luku-operaatioita on Log Analyticsin mukaan 350 kappaletta kuukaudessa ja niiden keskimääräinen koko on 3,3 mebitavua/3,5 megatavua. Luku-operaatiot kohdistuvat vain osaan tauluista. Kustannuksiin on lisättävä myös kirjoitus-operaatioksi laskettava palvelutason tyypin muutos (Microsoft 2023n).

Taulukko 1. Harvoin käytettävät, arkistoitavat tiedot

blobien määrä	tiedostokoko yht.	luku-operaatiot 30pv (lkm ja keskim.koko)	hot tier (säilytys) /kk	hot tier/ luku-operaatiot mukana	archive tier (säilytys) hinta/kk näillä luku-op. määrillä	kirjoitus-operaatiot/ hinta palvelutason muutoksesta
180	240 Tt = 219,15 TiB	350 kpl/3,5 Mt (3,3 MiB)	4 296,87 e	4 298,68 e	2 924,38 e	25,89 e (hot tier → archive tier)

Arvioidessani käyttötapauksen kustannuksia syötin Azuren kustannuslaskuriin ensin tarvittavan tallennuskapasiteetin koon. Olettaen, että tiedon määrä pysyy samana, tulee tallennustilaa olla 240 teratavua. Seuraavaksi lisäsin luku-operaatioiden määrän sille tarkoitettuun kohtaan laskurissa. Koska operaatiot laskutetaan 3.81 mebitavun/4 megatavun suuruisina osuuksina, tätä pienemmät luku-operaatioiden koot menevät ikään kuin ”hukkaan”. Käytin luku-operaatioiden kokona siis tuota minimiveloituksen arvoa (3,81 MiB/4 Mt).

Kuumalla palvelutasolla luku-operaatioista tuleva lisäkustannus varsinaisen tiedon säilytyksen päälle on suhteessa hyvin pieni, vain 1,81 euron suuruinen. Arkistointiin tarkoitetulla palvelutasolla jo yhden luku-operaation hinnaksi tulee yli 7 euroa. Tämän kertominen 350 luku-operaatiolla tuottaa yli 2500 euron kustannuksen, kun taas itse säilytyksen osuus on vajaat 408 euroa. Arkistointiin tarkoitetulla palvelutasolla säilytykseen tarvittava levytila on siis halpaa.

Kirjoitus-operaatiot voidaan Azuren kustannuslaskurin sijasta laskea myös seuraavan taulukon mukaisesti:

Transaction

	Premium	Hot	Cool	Cold	Archive
Write Operations* (every 4 MB, per 10,000)	€0.02729	€0.06472	€0.11984	€0.21571	€0.14381
Read Operations** (every 4 MB, per 10,000)	€0.00219	€0.00517	€0.01199	€0.11984	€7.19027
Query Acceleration - Data Scanned (per GB)	N/A	€0.00208	€0.00208	€0.00240	N/A
Query Acceleration - Data Returned (per GB)	N/A	€0.00074	€0.00922	€0.01199	N/A

Kuva 8. Kirjoitus-operaatioiden hintataulukko, jonka pohjalta tein arviot palvelutason muutoksesta syntyvistä kustannuksista (Microsoft s.a. b)

Valitsin tässä käyttötapauksessa tietojen siirtämiseksi palvelutasolta toiselle *Put Blob* -operaation, joka laskutetaan kirjoitus-operaatiotaulukon mukaan. Esimerkkitapauksessa kustannus otetaan taulukon *Archive write operation* -kohdasta. *Put Blob* -operaation laskutus perustuu siirrettävien blobien määrään. Kustannuslaskurin mukaan operaation alin laskutettava koko on 3,81 mebitavua/4 megatavua. (Microsoft 2023n). Esimerkkitapauksessa arkistointiin tarkoitetulle palvelutasolle siirtämisen hinnaksi tulee 0,14381 x 180 eli vähän päälle 25 euroa.

Githubista löytyy Microsoftin suosittelema kustannusten arviointia helpottava excel-työkirja, jonka välilehdillä voi arvioida erilaisia palvelutason valinnan tai arkistoinnin toteuttamisen skenaarioita (Microsoft 2023o). Tämä on erityisen hyödyllinen lisämateriaali, jos suunnitelmissa on arkistointiin tarkoitetulle palvelutasolle siirtäminen.

Mitä tulee ottaa huomioon, jos tiedot siirretään arkistointiin tarkoitetulle palvelutasolle?

Arkistointiin tarkoitetulle palvelutasolle siirtämisessä täytyy huomioida tietojen palauttamisen kesto, jos niitä tarvitaankin vielä aktiivisessa käytössä. Esimerkiksi 9,31 gibitavun/10 gigatavun kokoisen tiedoston saaminen uudelleen käyttöön saattaa kestää jopa 15 tuntia. Data pitää siirtää toiselle palvelutasolle, ennen kuin sitä pystytään käyttämään. Toiminnon nopeuttamisesta, *Archive high priority read/ retrieval*, syntyy lisäkustannuksia. Arkistointiin tarkoitettu palvelutaso on offline-tilainen, eikä sillä olevaa tietoa ole mahdollista muokata. Dataa tulee säilyttää vähintään 180 vuorokautta; aiemmasta tiedon poistamisesta tai siirtämisestä lämpimämmälle palvelutasolle aiheutuu ”rangaistusmaksuna” lisäkustannuksia. (Microsoft 2023d, Microsoft 2023n).

Tietojen siirtämisessä *archive tier* -palvelutasolle kustannuksia syntyy kirjoittamisesta (*write*), tiedon säilyttämisestä ja aiemmin mainitusta tiedon saattamisesta uudelleen käyttöön tarvittaessa (*rehydration*). Palvelutason muutos voidaan tehdä esimerkiksi REST API -rajapinnan kautta joko *Put Blob*, *Put Block*, *Put Block List* -operaatioilla tai AzCopy:n kautta *Set Blob Tier* -operaatiolla. Käytin esimerkkikustannusten laskemisessa *Put Blob* -vaihtoehtoa, jolloin palvelutasoissa ”alaspäin” siirtymisestä syntyi arkistoon kirjoittamisen (*write*) -kustannus. (Microsoft 2023n).

Arkistointiin tarkoitettu palvelutaso toimii kustannuksia säästävänä vaihtoehtona myös Azuren virtuaalikoneiden ja niillä pyörivien SQL-tietokantapalvelinten varmistusten säilytykseen. Yli kuusi kuukautta säilytettävät varmistukset kannattaa siirtää arkistointiin tarkoitetulle palvelutasolle esimerkiksi kuukausittain tai vuosittain. (Valiramani 2023, 22).

Microsoft suosittelee, että arkistointiin tarkoitetulle palvelutasolle siirrettäessä pienemmät tiedostot paketoidaan suuremmiksi TAR- tai ZIP-tiedostoiksi. Tämä mahdollistaa kustannussäästöt luku- ja

kirjoitusoperaatioiden määrän vähentämisen kautta, kun kutsuttavia kohteita on vähemmän. (Microsoft 2023p). Suuremmat tiedostot ylipäätään mahdollistavat paremman suorituskyvyn. Ideaali tiedostokoko ADLS Gen 2:ssa on 93-238 gibitavua/ 100-256 gigatavua. (Microsoft 2023q).

Säästääkö arkistointiin tarkoitetulle palvelutasolle siirtäminen kustannuksia esimerkkitapauksessa?

Kyllä, jos luku-operaatioiden määrään tai muuhun dataan kohdistuvaan käyttötarpeeseen ei tule huomattavia muutoksia.

4.2 Käyttötapaus 2: Staging-aineisto

Tämän käyttötapauksen dataa säilytetään sen varmistamiseksi, että siihen pystytään tarpeen vaatiessa palaamaan, jos dataa käyttävässä analytiikan sovelluksessa havaitaan virhe, ja data pitäisi ladata sinne uudestaan. Datalla ei ole pitkää säilytystarvetta, vaan se kulkee tilapäisen tallennuksen eli *staging*-alueen läpi matkallaan analytiikan sovelluksiin. Liiketoiminnan kanssa on sovittu, että dataa säilytetään vuoden ajan.

Esimerkkitapauksen dataa joko käytetään tai ei, joten kustannusten ennakoinnissa on varauduttava joko datan säilyttämiseen kuumalla palvelutasolla ”turhaan” tai toisaalta siihen, että siihen tuleekin viileällä palvelutasolla runsaasti luku-operaatioita. Arvioin näiden mahdollisten luku-operaatioiden määrän samaksi kuin blobien määrä.

Staging-alueen dataa on yhteensä 100,04 tebitavua/ 110 teratavua. Jos dataan ei kohdistu vuoden aikana lainkaan käyttöä luku-operaatioiden muodossa, sen säilyttäminen tulee ilman muuta edullisemmaksi viileällä palvelutasolla. Siinä tapauksessa, että dataa käytettäisiin ja luku-operaatioita syntyisi, lasketaan ennustetut kustannukset vuoden ajalle.

Taulukko 2. Staging-aineisto

blobien määrä	tiedosto-koko yht.	luku-ope-raatiot 30pv (lkm ja keskim. koko)	hot tier: pelkkä säily-tys kk ja vuosi	hot tier: säilytys ja luku-operaa-tiot (20 136 kpl) kk ja vuosi	cool tier:: pelkkä säilytys kk ja vuosi	cool tier: säilytys ja luku-operaatiot (20 136 kpl)	kirjoitus-operaatiot/hinta palvelutason muu-toksesta
20 136	110 Tt = 100,04 TiB	20 136 kpl/ 4 Mt (3,81 MiB)	1 989,85 e/kk x 12 =23 878,20e	2093,80 e/kk x 12 =25 125,60e	1038,35 e/kk x12 = 12 460,20 e	1279,65e/kk x 12 = 15 355,80 e	2 413,05 e (hot tier → cool tier)

Tein käyttötapauksen tietojen syöttämisen Azuren kustannuslaskuriin siten, että tarvittavaksi tallennuskapasiteetiksi tuli 100,04 tebitavua/110 teratavua ja luku-operaatioiden määräksi merkitsin joko nollan tai kaikkiin blobeihin kohdistuvan 20 136 kappaletta. Molemmilla vertailtavilla palvelutasoilla luku-operaatioiden aiheuttama lisäkustannus säilyttämisen päälle on suhteessa kokonaissummaan yllättävän vähäinen. Itse datan teratavumäärä on suuri, mutta luku-operaatioilla on optimaalinen 3,81 mebitavun/4 megatavun koko.

Kirjoitus-operaatiot kuumalta palvelutasolta viileälle siirrettäessä tehdään tässäkin esimerkissä *Put Blob* -komennolla ja niistä syntyvä kustannus on laskettava kuvan 8 taulukon mukaisesti viileän palvelutason hintojen päälle. Kustannus perustuu blobien määrään, sillä yhden blobin siirtäminen lasketaan yhdeksi operaatioksi.

Viileällä palvelutasolla kannattaa Microsotin mukaan säilyttää satunnaisesti käytettävää dataa. Minimisäilytysaika ilman ”rangaistusmaksua” viileällä palvelutasolla on 30 vuorokautta. Tiedonsiirtoviiveeseen viileän palvelutason säilytyksellä ei ole vaikutuksia. Saatavuudessa (*availability* eli tiedon käytettävissä oleminen) on kuumaan palvelutasoon verrattuna pieni ero: kun kuuman palvelutason saatavuudeksi luvataan 99,99%, on viileän palvelutason vastaava luku 99%. Vain-luku -varmennuksia (*RA-GRS*) käytettäessä saatavuusarvot ovat kuuman palvelutason kohdalla 99,99% ja viileällä palvelutasolla 99,9%.(Microsoft 2023d).

Tuleeko datan säilyttäminen edullisemmaksi kuumalla palvelutasolla vai viileällä palvelutasolla?

Data kannattaa siirtää viileälle palvelutasolle. Vaikka luku-operaatioita olisi näinkin runsaasti kuin esimerkissä, ero tiedon säilytyksessä kuuman ja viileän palvelutason välillä on loppujen lopuksi aika pieni. Lisäksi voi hyvinkin olla, ettei dataa tarvitsekaan käyttää (lukea) vuoden aikana.

4.3 Käyttötapaus 3: Massiivinen data, jota luetaan paljon

Tässä käyttötapauksessa on tarkastelun kohteena suuren, useita satoja teratavuja sisältävän datan säilyttäminen kuumalla palvelutasolla ja mahdollisuudet viileämmälle palvelutasolle siirtämiseen.

Laajoissa analytiikkaa palvelevissa tietoaltaissa dataa on usein huomattavia määriä ja ainakin osaa siitä sitä käytetään paljon. Rakensin esimerkin, jossa 181,90 tebitavun/ 200 teratavun aineisto säilytetään kokonaisuudessaan kuumalla palvelutasolla, ja siihen kohdistuu kymmeniä miljoonia luku-operaatioita kuukausittain. Oletuksena on, että koko datalla ei kuitenkaan todellisuudessa ole pitkäaikaista kuumalla palvelutasolla säilyttämisen tarvetta.

Taulukko 3. Massiivinen data, jota luetaan paljon

blobien määrä	tiedostokoko yht.	luku-operaa-tiot 30 pv (lkm ja kes-kim. koko)	hot tier (pelkkä säilytys) /kk	hot tier/ luku-operaatiot mukana	cool tier/ luku-operaa-tiot	cool tier (pelkkä säily-tys) hinta/kk	kirjoitus-operaatiot/hinta palvelutason muutoksesta
8 000 000	200 Tt = 181,90 TiB	49 000 264 kpl/ 1,9 Mt (1,81 MiB)	3 587,02 e	256 538,24e	589 096,08 e	1 887,91 e	958 702,06 e (hot tier → cool tier)

Tein kustannusarvion Azuren laskurilla samaan tapaan kuin aiemmissa käyttötapauksissa. Keskityn tässä esimerkissä etsimään keinoja kustannusten säästämiseksi dataan kohdistettavien hallintatoimien kautta. Näillä luku-operaatioiden määrillä dataa ei kokonaisuudessaan kannata siirtää pois kuumalta palvelutasolta. *Put Blob* -komennolla toteutettavan palvelutason muutoksen ja siitä johtuvien, blobien määrään perustuvien kirjoitus-operaatioiden hinnaksikin tulee lähes miljoona euroa.

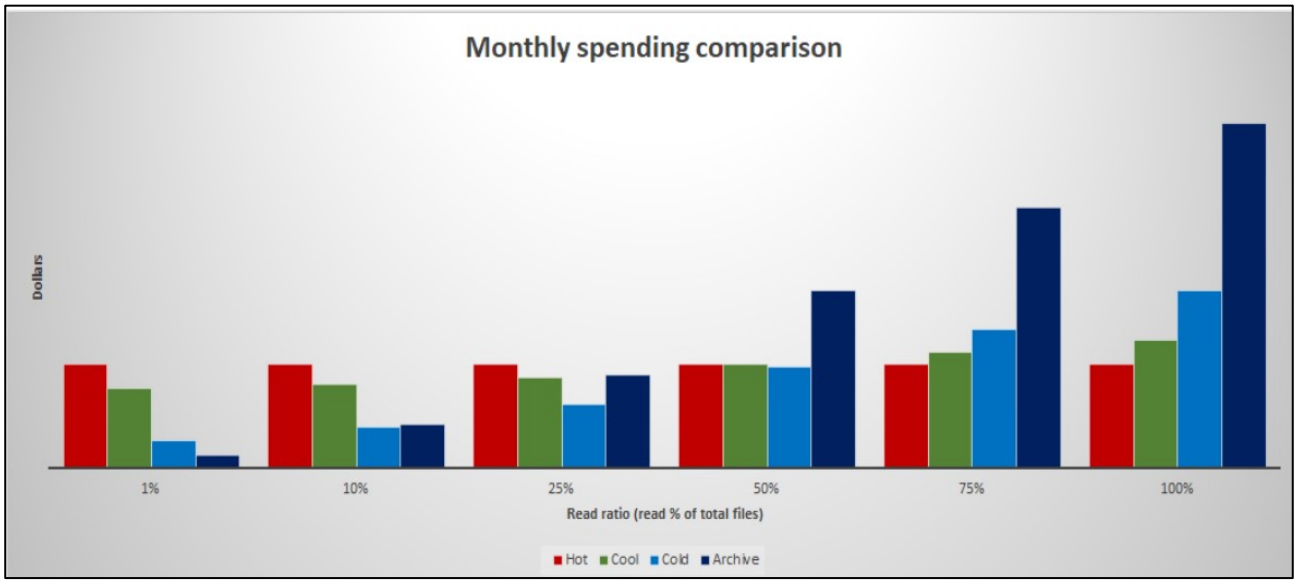
Mitä keinoja on käytettävissä kustannussäästöjen löytämiseksi tässä tapauksessa?

Kustannussäästöjen saavuttamiseksi datan käyttö on hyvä kartoittaa esimerkiksi skeemoittain tai kansioittain. Tämä on mahdollista tehdä Log Analyticsin avulla, käyttölokeja tutkimalla. Jos datasta löytyy harvemmin käytettäviä tiedostoja, ne kannattaa siirtää säilytettäväksi käyttötiheyden ja -tarpeen mukaisesti viileälle, kylmälle tai arkistointiin tarkoitetulle palvelutasolle. Dataan kohdistuvat luku-operaatiot kannattaa myös jakaa prosentuaalisiin osuuksiin, kuten Microsoft suosittelee (Microsoft 2023p). Kannattaa myös varmistaa, ettei datassa ole mukana duplikaatteja ja jos on, poistaa turhat kaksoiskappaleet säilytyskustannusten pienentämiseksi (Guruswamy 2024).

Yksi keino hallita datan kustannuksia on Azuren Lifecycle Management -työkalun käyttö. Vaikka on selvää, että Lifecycle Management-työkalun käyttö helpottaa kustannusten hallintaa, sen tehokas hyödyntäminen vaatii etukäteissuunnittelua. Datan käyttöä on pystyttävä jonkin verran ennakoimaan, jotta siirtäminen palvelutasolta toiselle ei itsessään aiheuta turhia kustannuksia. Esimerkiksi jos data siirtyy elinkaarenhallinnan säännön mukaan kuumalta arkistointiin tarkoitetulle palvelutasolle 30 päivän kuluttua sen luontihetkestä, mutta dataa tarvitaankin puolen vuoden kuluttua aktiivisesti, saattaisi jopa olla edullisempaa pitää se kuumalla palvelutasolla. Dataan kannattaa siis tutustua.

Käyttötarpeessa on hyvä huomioida tiedonhakuviiveen merkitys. Siinä missä kuumalla, viileällä ja kylmällä palvelutasolla data on saatavilla käytännössä reaaliaikaisesti (millisekunneissa), arkistointiin tarkoitetulla palvelutasolla data on käytettävissä vasta, kun se on ensin siirretty sieltä lämpimämmälle palvelutasolle, jolloin puhutaan jopa 15 tunnista. Lisäksi datan nouto ja lukeminen toiselle palvelutasolle maksavat. (Microsoft 2023n).

Sopivaa palvelutasoa voi yrittää hahmottaa arvioimalla vaikka karkeastikin, kuinka suurta prosenttiosuutta datasta luetaan kuukausittain. (Microsoft 2023p).



Kuva 9. Kun datan määrästä yli puolet on kuukausittain luku-operaatioiden kohteena, ei arkistointiin tarkoitettu palvelutaso missään tapauksessa ole kustannustehokas tapa säilyttää sitä (Microsoft 2023p)

Microsoftin kustannuslaskurin nuolivalitsimen avulla on mahdollista haarukoida, mikä on se täsmällinen luku-operaatioiden määrä, joka muuttaisi viileällä palvelutasolla säilyttämisen edullisemmaksi kuin kuumalla palvelutasolla. Jos lähes 50 miljoonan luku-operaation määrää saataisiin esimerkkitapauksessa vähennettyä 21 249 590:een tai sen alle, viileällä tasolla säilyttäminen tulisi edullisemmaksi. Tämän lisäksi toki tulee huomioida palvelutason muutoksesta aiheutuvat kirjoitus-operaatioiden kustannukset. Näin suurella blobien määrällä, olettaen että kaikki data siirrettäisiin viileälle palvelutasolle, kustannukset palvelutason vaihdosta ovat huomattavat. Joka tapauksessa luku-operaatioiden määrän vähentämiseen kannattaa mahdollisuuksien mukaan pyrkiä.

Edellä mainittujen, datan palvelutasoon suoraan liittyvien kustannusten hallinnan lisäksi Microsoft suosittelee, että tietoaltaassa säilytettävän datan tiedostomuodot, tiedostokoot ja kansiorakenne suunnitellaan kustannustehokkaaksi jo ennalta. Erityisesti analytiikan tarpeisiin suositellaan

Apache Parquet -tiedostomuotoa, sillä se mahdollistaa vain tarpeellisten tietosarakkeiden lukemisen operaatioissa. Kuten aiemmin jo havaittiin, tiedostojen suurempi koko on myös hyödyksi. (Microsoft 2023q).

Suurta datamäärää voidaan myös koettaa hallita indeksoinnin avulla. Blobeihin voidaan lisätä “tageja”, jossa niiden sisältöä kuvataan avaimella ja arvolla (*key-value pair*), kuten esimerkiksi viimeisimpään käsittelyyn viittaavalla päivämäärällä. Tallennustilillä jo olevien blobien indeksointi tapahtuu *SetBlobTags* -rajapintakutsulla. ”Tagien” arvoja voidaan käyttää rakennettaessa Azuren Lifecycle Management -työkalulla sääntöjä siitä, milloin blobien siirtäminen sopivimmalle palvelutasolle on mahdollista. (Microsoft 2023r).

Jo tietoaltaan suunnitteluvaiheessa olisi hyvä tietää, minkä tyyppistä dataa tullaan tallentamaan, miten data viedään tietoaltaaseen, kuka käyttää dataa ja miten sitä käytetään (Microsoft 2022). Toisaalta, jos organisaatiolla ei ole aiempaa kokemusta tietoaltaasta analytiikan käytössä, ennustaminen voi olla hyvin vaikeaa. Kuitenkin datan vieminen suoraan optimaaliselle palvelutasolle säästäisi kustannuksia ja työtä.

5 Tulokset ja johtopäätökset

Kävin opinnäytetyötäni varten läpi runsaasti varsinkin Microsoftin tarjoamaa lähdeaineistoa. Kirjallisuuskatsauksen keinoin kerätty ja analysoitu, Azure Data Lake Storage Gen 2 -tietoaltaan säilytyskustannuksiin liittyvä tietoperusta toimi pohjana itse keksimilleni käyttötapauksille. Näiden käyttötapausten avulla pystyin tutustumaan kustannuksiin vaikuttaviin tekijöihin konkreettisemmin kuin pelkkää lähdeaineistoa tulkitsemalla. Tässä opinnäytetyöni viimeisessä luvussa teen yhteenvedon lähdeaineiston hyödyntämisen ja käyttötapausten tutkimisen yhdistelmällä löytämistäni vastauksista tutkimuskysymyksiini, teen johtopäätöksiä tutkimukseni aiheesta ja kuvaan oppimisprosessiani.

5.1 Tulosten tarkastelu

Tutkimuskysymyksilläni halusin selvittää,

> *Miten saadaan selville Microsoft Azure Data Lake Storage Gen 2 -palvelun säilytyskustannusten arvioinnissa tarvittavat lukumäärätiedot luku-operaatioille ja tiedostojen koolle?*
>
> *Minkälaisessa tilanteessa datan siirtäminen kuumalta palvelutasolta viileämmälle palvelutasolle tuottaa kustannussäästöä?*
>
> *Minkälaisia rajoituksia viileämmällä palvelutasolla säilyttämiseen liittyy ja mitä tulee ottaa huomioon siirrettäessä dataa palvelutasojen välillä?*
>
> *Miten palvelutasoon liittyvää kustannusoptimointia voidaan käytännössä edistää ja toteuttaa?*

Esittelin opinnäytetyössäni Azure Storage Explorerin ja Azure Log Analyticsin työvälineinä, joista on apua etsittäessä blobeihin kohdistuvien luku-operaatioiden määriä ja blobien keskimääräisiä tiedostokokoja. Nämä tiedot, samoin kuin arvio tallennustilillä tarvittavasta kokonaistallennuskapasiteetista, tarvittiin Azuren kustannuslaskuria varten. Kustannuslaskuriin syötettyjen tietojen avulla oli mahdollista saada käsitys edullisimmasta palvelutasosta.

Ensimmäisessä käyttötapauksessani datan siirtäminen arkistointiin tarkoitetulle palvelutasolle osoittautui kannattavaksi. Dataan kohdistuvat luku-operaatiot eivät olleet määrältään niin merkittäviä, että arkistointiin tarkoitetulla palvelutasolla säilyttäminen olisi tullut kannattamattomaksi verrattuna kuumaan palvelutasoon. Tässä tapauksessa kuitenkin oli huomioitava ennen kaikkea se, että

datan palauttaminen arkistointiin tarkoitetulta palvelutasolta kestäisi jopa tunteja ja tuottaisi lisäkustannuksia.

Toisessa käyttötapauksessani blobeihin kohdistuvien luku-operaatioiden koko (3,81 mebitavua/4 megatavua) osoittautui kustannusnäkökulmasta optimaaliseksi. Koska luku-operaatiot prosessoidaan ADLS Gen2 -tietoaltaassa tähän kokoon ”pilkottuina”, ei niiden kohdalla tässä käyttötapauksessa tarvinnut maksaa tyhjästä. Luku-operaatiot laskutetaan joka tapauksessa edellä mainitun koon mukaan, vaikka ne olisivat pienempiä.

Toisen käyttötapauksen kustannuslaskelmia tehdessäni havaitsin, että kuuman palvelutason ja viileän palvelutason väliset erot hinnoissa ovat yllättävän pienet. Luku-operaatioiden määrä viileällä palvelutasolla on suhteessa datan kokoon eli säilytettäviin tavumääriin verrattuna epäolennaisempi kuin ennalta ajattelin.

Kolmannessa käyttötapauksessa luku-operaatioiden määrä oli valtava ja niiden koko epäedullinen. Tämänkaltaisen tapauksen selvitteleminen vaati tutustumista erilaisiin keinoihin, joita Microsoft ja muut lähteet suosittelevat kustannusten hallintaan. Keinoista nousivat esille datan jakaminen käytön mukaisiin kokonaisuuksiin vaikkapa prosentuaalisesti arvioituna; käytännössä sen kansiointi ja luokitteleminen eri palvelutasoille tämän mukaisesti tiedonhakuviiveet huomioiden, duplikaattien poisto, Azuren Lifecycle Management -työkalun käyttö viimeisimpään tiedon käyttöön perustuvasti, datan indeksointi ja erityisesti analytiikan käyttöön soveltuvat tallennusmuodot. Käsitys datan käytöstä on mahdollista perustaa lokitietojen tutkimiseen. Tämän käyttötapauksen yhteydessä havaitsin, että datan siirtäminen eri palvelutasojen välillä on suurilla blobien määrillä hyvin kallista.

Tulin kaikkien käyttötapausten kohdalla tavalla tai toisella todistaneeksi Microsoftin hinnoitteluperiaatteen, joka lähtee siitä, että datan säilyttäminen kuumalla palvelutasolla on kallista, ja viileämmillä tasoilla halpaa – mutta datan käyttö varsinkin arkistointiin tarkoitetulla palvelutasolla on kallista. Tiedonhakuviiveen merkitys muilla kuin arkistointiin tarkoitetulla palvelutasolla jää käytännössä olemattomaksi. Datan käytettävissä olemisen (saavutettavuuden) palvelulupauksessa on minimaalisia eroja eri palvelutasojen välillä.

5.2 Johtopäätökset

Ideaalitilanteessa datan palvelutaso on kustannustehokasta päättää etukäteen ja ladata data suoraan toivotulle palvelutasolle. Tällöin ei tarvitse maksaa sekä datan alkuperäisestä lataamisesta

että sen päälle palvelutason muutoksesta syntyvistä kirjoitus-operaatioista. (Microsoft 2023p). Toisaalta voi olla vaikea tietää etukäteen, miten dataa tullaan käyttämään ja palvelutasoa koskeva päätös voidaan joka tapauksessa joutua pyörtämään. Azuren Lifecycle Management -työkalun käyttö vaatii niin ikään ennakkosuunnittelua, jotta sillä luodut säännöt datan siirtämisestä palvelutasolta toiselle eivät käänny itseään vastaan. Tämänkaltaisiin suunnittelutehtäviin kannattaa kustannusoptimoinnin nimissä organisaatioissa panostaa.

Analytiikan tarpeisiin kerättävää dataa kertyy usein suuria määriä. Käytännössä karkeat arviotkin siitä, miten data ”käyttäytyy” ovat avuksi kustannustehokkainta säilytyksen palvelutasoa valittaessa. Microsoftin suosittelema arvion tekeminen dataan kohdistuvien luku-operaatioiden prosentuaalisista osuuksista olisi jo isoksi avuksi. Tällaisen arvion tekeminen käytännössä jää Microsoftin aineistossa avoimeksi. Lokitietoja analysoimalla arviointi olisi mahdollista, mutta tässä vaiheessa aloin jo toivoa kertasilmäyksellä saatavaa kuvaa datan käytöstä jollakin monitorointivälineellä toteutettuna.

Tiedostokokoja tai tarvittavaa tallennuskapasiteettia kuvaavina yksiköinä binääriset ja metriset yksiköt vaativat tarkkuutta opinnäytetyötä toteuttaessani. Microsoftin kustannuslaskuri käyttää metrisiä yksiköitä eli esimerkiksi gigatavuja, kun taas ADLS Gen 2-tietoaltaaseen tallennetun tiedon tarkasteluun käyttämäni Azure Storage Explorer käyttää binäärisiä muotoja, kuten tebitavuja. Tiedon tallennus luultavasti oikeasti tapahtuu binääristen yksiköiden mukaan, onhan se tietokoneen käyttämä tapa. Olisi ollut kätevää, jos Microsoftin kustannuslaskurissakin olisi huomioitu tiedon tallennus binäärisenä. Datan määrän kasvaessa 2,4 prosentin ero ”mittayksikössä” alkaa kuitenkin olla joissain tapauksissa merkittävä.

Säilytyksen palvelutasoihin tai niiden hinnoitteluun kriittisesti suhtautuvaa lähdeaineistoa oli vaikea löytää. Kaipasin lähteitä, joissa olisi käytännön kokemusten perusteella tehty suosituksia siitä, miten kustannusoptimointi olisi mahdollista ja mahdollisimman helppoa saavuttaa. Pilvioperaattorit jättävät kustannusoptimoinnin organisaation itsensä varaan. Olisin odottanut löytäväni internetistä lähteitä, joissa organisaatiot jakaisivat kokemuksiaan säilytyksen palvelutasojen tai yleisemminkin pilveen toteutettujen tietoaltaiden kustannusten hallinnasta. Alalla olisi varmasti tarvetta verkostoille, joissa kokemuksia voisi jakaa.

Tutkimukseni luotettavuutta arvioitaessa on mainittava, että käyttötapauksissani huomioidaan vain osan kaikista Azuren säilytyskustannuksista. Tietoaltaan perustamiseen ja ylläpitoon, datan liikkumiseen ja esimerkiksi tietoturvasta huolehtimiseen vaaditaan lukuisia maksullisia komponentteja ja palveluita, joita en ole tässä opinnäytetyössä ottanut huomioon. Laskelmatkin ovat parhaimmillaan-

kin suuntaa-antavia, ja niissä voi olla virheitä. Hierarkkiseen nimiavaruuteen liittyvät metatietokustannukset jätin pois käyttötapausten laskelmista, vaikka todellisuudessa niistäkin syntyisi kuluja. Olennaisena pidän kuitenkin sitä, että pystyn esimerkkieni avulla konkretisoimaan tiedon säilytyksen palvelutasoja ja niiden ominaispiirteitä ADLS Gen 2 -tietoaltaassa.

5.3 Opinnäytetyöprosessin ja oman oppimisen arviointi

Microsoftin hinnoittelumallin ymmärtäminen oli haastavaa. Eniten vaivaa tuotti kustannusten laskemisen ymmärtäminen oikein ja lukujen sijoittaminen oikeille paikoilleen laskurissa. Kustannuslaskurin yhteyteen olisin kaivannut lisäinformaatiota infopainikkeiden taakse, jotta siihen osaisi helpommin syöttää oikeat luvut oikeisiin kohtiin. Hinnoittelussa tuntui olevan vaikka mitä huomioitavaa ja kokonaisuus monimutkainen. Kun olin lähdeaineiston tutkimisen ja saamani ohjauksen avulla ymmärtänyt, mitä lukuja tarvitsen, itse luku-operaatioiden määrien löytäminen oli oikeanlaisella KQL-lauseella Log Analytics-palvelussa melko helppoa. Azure Storage Explorerin käyttöliittymä oli helppokäyttöinen ja siihen liittyvää ohjemateriaalia saatavilla hyvin.

Opinnäytetyössäni jouduin pohtimaan, miten saan aiheen pidettyä rajattuna, sillä aineistoa varsinkin Microsoftin asiakassivuilla on erittäin runsaasti. Microsoftin lähteet ovat jakautuneet usealle eri verkkosivulle, ja niiden välillä poukkoilu on tarkkaavaisuutta vaativaa. Lähteille ominaista on myös niiden jatkuva päivittyminen. Pilvikomponenttien kehitys on niin nopeaa, että Microsoftin materiaalit saattavat vaatia ajantasaistamista muutaman kuukauden välein. Samasta syystä erilaiset blogilähteet ja kirjallisuus tuntuvat ”laahaavan perässä”, ja niissä kerrottujen tietojen oikeellisuus piti lopulta varmistaa uusimmista Microsoftin lähteistä.

Menetelmällisen kehyksen luominen vaati tutustumista useisiin tutkimussuuntauksiin. Opinnäytteeni ei ole suoraan sijoitettavissa yhteen lokeroon, vaan siinä on elementtejä ainakin kahdesta erilaisesta menetelmästä. Tutustuin opinnäytetyön ohjaajan ehdotuksesta ensimmäistä kertaa Chat GPT:n tarjoamiin mahdollisuuksiin opiskelun tukena. Käytin tämän tekoälysovelluksen ilmaisversiota joidenkin yksityiskohtien tarkastamiseen ja tutkimusmenetelmään liittyvään ideointiin. Microsoftin palvelutasojen ja kustannusten kohdalla Chat GPT:n tiedot olivat kuitenkin vanhentuneita.

Opinnäytetyöprosessi oli sujuva ja sain paljon hyödyllistä ohjausta. Vaikeinta ja eniten aikaa vievää oli lähteiden etsiminen ja suodattaminen, termien kääntäminen suomeksi, sekä Microsoftin kustannuslaskurin käytön opetteleminen. Pääsin monesta hankalasta hetkestä kuitenkin lopulta yli ja sain itseluottamusta opinnäytteen tekemisestä kokopäivätyön ohessa. Tulin kuin huomaamatta omaksuneeksi paljon tietoa, jota voin toivottavasti hyödyntää työelämässä.

Lähteet

Amazon Web Services 2024. What is Compute? Luettavissa: https://aws.amazon.com/what-is/compute/#:~:text=In%20cloud%20computing%2C%20the%20term,computational%20success%20of%20any%20program. Luettu: 6.5.2024.

Ari Hovi Oy. 2023. Data-alan termien selitykset ja kuvaukset. Blogiteksti. Luettavissa: https://www.arihovi.com/materiaalit/datapedia-data-alan-termit-avattuna/. Luettu: 27.3.2024.

Ari Hovi Oy. 1.12.2020. Tietoallas vai tietovarasto? Blogiteksti. Luettavissa: https://www.arihovi.com/tietoallas-vai-tietovarasto/. Luettu: 27.3.2024.

Ari Hovi Oy. 2.11.2021. Data Lakehouse – Tietovarasto ja tietoallas yhdessä. Blogiteksti. https://www.arihovi.com/data-lakehouse-tietovarasto-ja-tietoallas-yhdessa/. Luettu: 27.3.2024.

Azure Storage 2023. Azure Data Lake Storage Gen2 Billing FAQs. Luettavissa: https://azure.github.io/Storage/docs/analytics/azure-storage-data-lake-gen2-billing-faq/.Luettu: 27.3.2024.

Borgini, J. 2023. A short guide to Azure Data Lake Storage pricing. TechTarget. Luettavissa: https://www.techtarget.com/searchstorage/tip/A-short-guide-to-Azure-Data-Lake-Storage-pricing. Luettu: 28.3.2024.

Cheshire, J. 2022. Exam Ref AZ-900 Microsoft Azure Fundamentals. 3rd Edition. Microsoft Press. E-kirja. Luettu: 27.3.2024.

Donnelly, J. 21.9.2021. GB vs GiB: What's the Difference Between Gigabytes and Gibibytes? Luettavissa: https://massive.io/file-transfer/gb-vs-gib-whats-the-difference/. Luettu: 27.3.2024.

Dremio 2024. Read and Write Operations. Luettavissa: https://www.dremio.com/wiki/read-and-write-operations/. Luettu: 27.3.2024.

Gbmb.org 2024. GiB to GB Conversion. Luettavissa: https://www.gbmb.org/gib-to-gb. Luettu: 27.3.2024.

Gillespie, M. & Givre, C. 2021. Understanding Log Analytics at Scale. 2nd Edition. O'Reilly Media, Inc. E-kirja. Luettu: 27.3.2024.

Guruswamy, D. 2024. Azure Storage cost optimization to achieve maximum cost savings. Luettavissa: https://www.serverless360.com/blog/azure-storage-cost-optimization. Luettu: 28.3.2024.

Güntert, M. 22.1.2023. How to view and change access tiers for blobs and file shares of your Azure storage account. Blogiteksti. Luettavissa: https://www.azureblue.io/how-to-view-and-change-access-tiers-for-blobs-and-file-shares-of-your-azure-storage-account/. Luettu: 27.3.2024.

Hargreaves, B. & Zaal, S. 2020. Implementing Microsoft Azure Architect Technologies: AZ-303 Exam Prep and Beyond. Second Edition. Birmingham: Packt Publishing Ltd. E-kirja. Luettu: 27.3.2024.

Junni, S. 2020. Tietovarastoinnin hyvät käytänteet ja niiden toteutuminen tietovarastoprojektissa. AMK-opinnäytetyö. Laurea Ammattikorkeakoulu. Luettavissa: https://www.theseus.fi/bitstream/handle/10024/354329/Opinn%C3%A4ytety%C3%B6%20Saku%20Junni.pdf?sequence=2&isAllowed=y. Luettu: 27.3.2024.

Mannila, M. 11.2.2021. Kirjallisuuskatsaus opinnäytetyön muotona. Energiaa: Vaasan ammattikorkeakoulun verkkolehti. Luettavissa: http://urn.fi/URN:NBN:fife202102114568. Luettu: 27.3.2024.

Mansouri, Y. & Erradi, A. 2018. Cost Optimization Algorithms for Hot and Cool Tiers Cloud Storage Services. 2018 IEEE 11th International Conference on Cloud Computing. s. 622-629. Luettavissa: https://ieeexplore.ieee.org/document/8457855. Luettu: 27.3.2024.

Mansouri, Y. & Erradi, A. 2020. Online cost optimization algorithms for tiered cloud storage services. Journal of Systems and Software, Volume 160, February 2020. Luettavissa: https://www.sciencedirect.com/science/article/abs/pii/S0164121219302316?via%3Dihub. Luettu: 27.3.2024.

Microsoft. s.a. a. Data warehousing and analytics. Luettavissa: https://learn.microsoft.com/en-us/azure/architecture/example-scenario/data/data-warehouse. Luettu: 27.3.2024.

Microsoft. s.a. b. Azure Data Lake Storage pricing. Luettavissa: https://azure.microsoft.com/en-gb/pricing/details/storage/data-lake/#pricing. Luettu: 27.3.2024.

Microsoft 2022. Overview of Azure Data Lake Storage for cloud-scale analytics. Luettavissa: https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/best-practices/data-lake-overview?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json&bc=%2Fazure%2Fstorage%2Fblobs%2Fbreadcrumb%2Ftoc.json. Luettu: 24.4.2024.

Microsoft 2023a. Plan and manage costs for Azure Blob Storage. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/common/storage-plan-manage-costs. Luettu: 27.3.2024.

Microsoft 2023b. Azure Storage Explorer. Luettavissa: https://azure.microsoft.com/en-us/products/storage/storage-explorer. Luettu: 27.3.2024.

Microsoft 2023c. Optimize costs by automatically managing the data lifecycle. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview. Luettu: 27.3.2024.

Microsoft 2023d. Access tiers for blob data. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview. Luettu: 27.3.2024.

Microsoft 2023e. Introduction to Azure Data Lake Storage Gen 2. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction . Luettu: 27.3.2024.

Microsoft 2023f. Log Analytics tutorial. Luettavissa: https://learn.microsoft.com/en-us/azure/azure-monitor/logs/log-analytics-tutorial. Luettu: 27.3.2024.

Microsoft 2023g. Introduction to Azure Data Lake Storage Gen2. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction. Luettu: 27.3.2024.

Microsoft 2023h. Manage blob containers using the Azure portal. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/blob-containers-portal Luettu: 27.3.2024.

Microsoft 2023i. Storage account overview. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/common/storage-account-overview. Luettu: 27.3.2024.

Microsoft 2023j. Blob rehydration from the archive tier. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/archive-rehydrate-overview. Luettu: 27.3.2024.

Microsoft 2023k. Blob Snapshots. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/snapshots-overview. Luettu: 27.3.2024.

Microsoft 2023l. Set a blob's access tier. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-online-manage?tabs=azure-portal. Luettu: 27.3.2024.

Microsoft 2023m. Configure a lifecycle management policy. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-policy-configure?tabs=azure-portal. Luettu: 27.3.2024.

Microsoft 2023n. Estimate the cost of archiving data. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/archive-cost-estimation. Luettu: 27.3.2024.

Microsoft 2023o. Azure archive storage cost estimation. Luettavissa: https://azure.github.io/Storage/docs/backup-and-archive/azure-archive-storage-cost-estimation/azure-archive-storage-cost-estimation.xlsx . Luettu 27.3.2024.

Microsoft 2023p. Best practices for using blob access tiers. Luettavissa: https://learn.microsoft.com/en-gb/azure/storage/blobs/access-tiers-best-practices. Luettu: 27.3.2024.

Microsoft 2023q. Best practices for using Azure Data Lake Storage Gen2. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices. Luettu: 27.3.2024.

Microsoft 2023r. Manage and find Azure Blob data with blob index tags. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/blobs/storage-manage-find-blobs?tabs=azure-portal. Luettu: 21.4.2024.

Microsoft 2024a. Azure Data Lake Storage pricing. Luettavissa: https://azure.microsoft.com/en-au/pricing/details/storage/data-lake/#pricing. Luettu: 27.3.2024.

Microsoft 2024b. Microsoft Pricing Calculator. Luettavissa: https://azure.microsoft.com/en-us/pricing/calculator/. Luettu: 27.3.2024.

Microsoft 2024c. Get started with Storage Explorer. Luettavissa: https://learn.microsoft.com/en-us/azure/storage/storage-explorer/vs-azure-tools-storage-manage-with-storage-explorer?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json&bc=%2Fazure%2Fstorage%2Fblobs%2Fbreadcrumb%2Ftoc.json&tabs=windows. Luettu: 27.3.2024.

Microsoft 2024d. Storage Blob Logs. Luettavissa: https://learn.microsoft.com/en-us/azure/azure-monitor/reference/tables/storagebloblogs. Luettu: 27.3.2024.

Puntanen V. & Virkkunen, M. 5.3.2024. Microsoftin asiantuntijat. Opinnäytetyöhön liittyviä kysymyksiä (kustannukset). Sähköpostiviestit.

Pure Storage .2024. Blob Storage vs. File Storage. Blogiteksti. Luettavissa: https://blog.purestorage.com/purely-informational/blob-storage-vs-file-storage/. Luettu: 27.3.2024.

Reis, J. & Housley, M. 2022. Fundamentals of data engineering: plan and build robust data systems. O'Reilly Media, Inc. First edition.

Salesforce Inc. 2024. Business Intelligence vs. Business Analytics: What's The Difference? Luettavissa: https://www.tableau.com/learn/articles/business-intelligence/bi-business-analytics. Luettu: 3.4.2024.

Salminen, A. 2011. Mikä kirjallisuuskatsaus? Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin. Vaasan yliopiston julkaisuja. Opetusjulkaisuja 62: Julkisjohtaminen 4. Vaasa. Luettavissa: https://www.uwasa.fi/materiaali/pdf/isbn_978-952-476-349-3.pdf. Luettu: 28.3.2024.

Slingerland, C. 19.4.2023. The No BS Guide To Understanding Azure Storage Costs. Luettavissa: https://www.cloudzero.com/blog/azure-storage-costs/. Luettu: 28.3.2024.

Tietoevry. 5.5.2020. Ota datamaailman termit haltuun datasanaston avulla. Blogeja teknologiasta ja liiketoiminnan muutoksesta. Luettavissa: https://www.tietoevry.com/fi/uutishuone/kaikki-uutiset-ja-tiedotteet/blogi/2020/ota-datamaailman-termit-haltuun-datasanaston-avulla/#:~:text=Dataputki%20(data%20pipeline)%20on%20hallittu,ennuste%2C%20jota%20k%C3%A4ytet%C3%A4%C3%A4n%20rajapinnan%20kautta. Luettu: 27.3.2024.

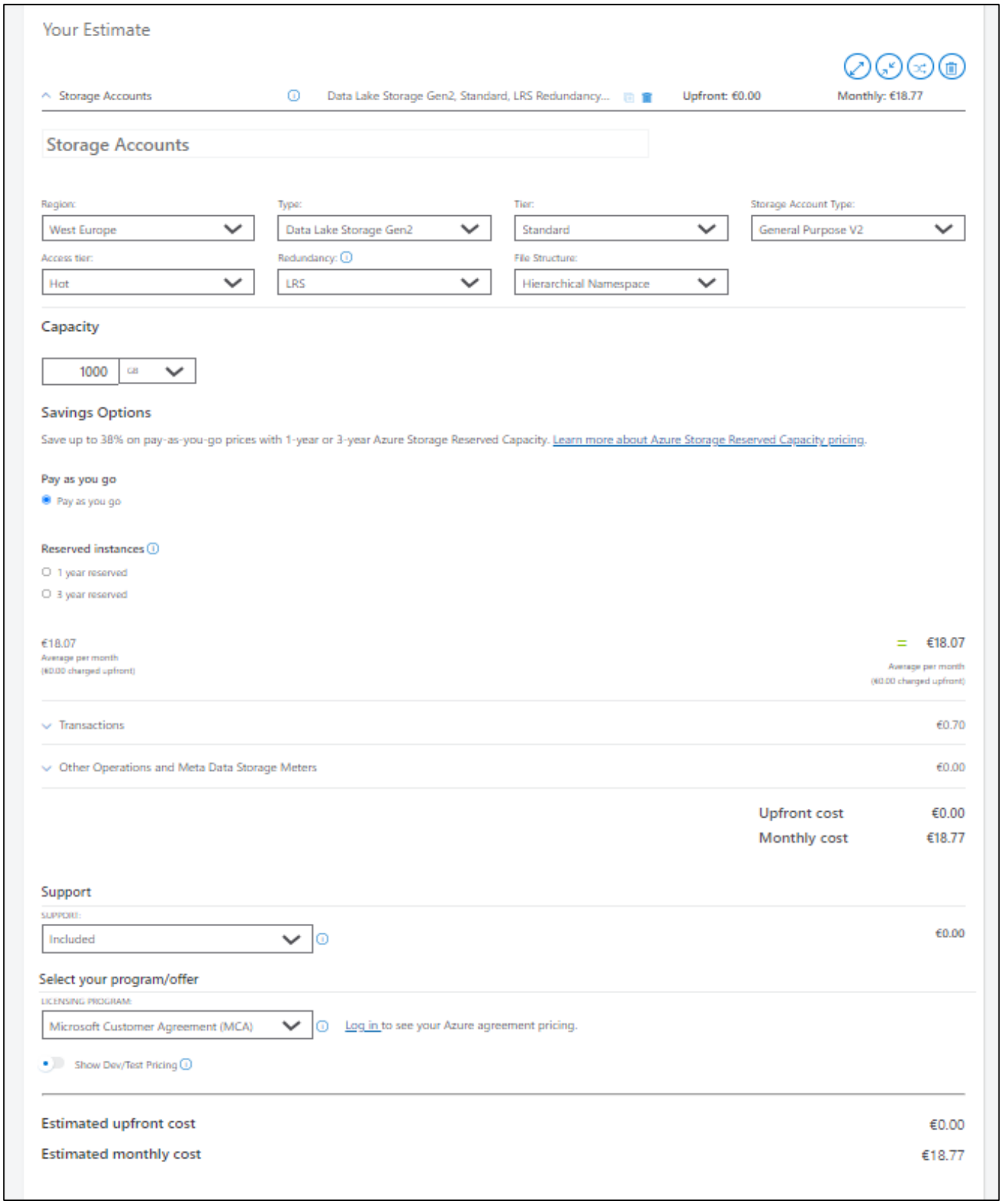
Törmänen A. 2017. Johdanto tietovarastointiin. CreateSpace.

University of Babylon 2018. College of Computer Technology. Basic Memory Operations. Luettavissa: https://www.uobabylon.edu.iq/eprints/publication_3_2538_1575.pdf. Luettu: 27.3.2024.

Valiramani, A. 2023. Microsoft Azure monitoring & management: the definitive guide. Pearson Education, Inc.

Veritas Technologies. 28.9.2022. When not to use Azure Storage Reserved Capacity blog post. Blogiteksti. Luettavissa: https://www.veritas.com/blogs/when-not-to-use-azure-storage-reserved-capacity-blog-post. Luettu: 27.3.2024.

Liite 1. Microsoftin tarjoama kustannuslaskuri



Kustannuslaskuri keväällä 2024 (Microsoft 2024b).