Teemu Pöytäniemi

Tiedostojen siirto ja talletus Azure -pilvialustalla

Diplomityö

Tieto- ja säjkötekniikan tiedekunta

Tarkastaja: Kari Systa

9/202

TIIVISTELMÄ

Teemu Pöytäniemi: Tiedostojen siirto ja talletus Azure -pilvialustalla

Diplomityö

Tampereen yliopisto

Tietotekniikan tutkinto-ohjelma

9/2024

Diplomityössä tarkastellaan tiedostojen siirtämistä ja tallentamista Azure -pilvialustalla. Tiedostojen siirtämistä ja talletusta tarkastellaan kolmesta näkökulmasta: kustannukset, suorituskyky ja tietoturva. Tiedoston siirtämistä vertaillaan eri tiedostonsiirtoprotokollien avulla. Diplotyössä vertaillaan seuraavia suojaamattomia protokollia: HTTP, WebDav, WebSocket sekä FTP. Lisäksi työssä on mukana kaikkien suojaamattomien prokollien suojatut versiot sekä suojattu SFTP- protokolla. Tiedostojen talletukseen käytetään SQL-tietokantaa, NOSQL-tietokantana, Object Storagea, Block Storagea ja File Storagea. Tietoturvan teoriaosuudessa käsitellään RSA- ja TLS- salausta.

Diplotyötä varten pystytettiin testausympäristö Azureen. Azuren virtuaalikoneeseen pystytettin SFTP-, FTPS- ja WebDav -palvelimet tiedonsiirtoa varten. .NET -palvelimen pystytettiin HTTP- ja WebSocket- liikennettä varten. Angularilla tehdyn käyttöliittymän avulla pystyttiin testaamaan HTTP- ja WebSocket liikennettä selaimesta. C#- ja Python -kirjastoja käytettiin testaamaan tiedonsiirtoa eri protokollilla. Tiedostojen tallentamiseen käytettiin Postgres Tietokantaa, Cosmos- tietokantaa, Azure Blob -tietovarastoa, Azure Files- tietovarastoa sekä Azure managed disks- tietovarastoa.

Kustannusvertailussa vertaillaan Azuren talletusratkaisuiden hintoja kolmen eri talletuskoon avulla: 16GB , 245GB ja 1 TB. Talletusratkaisuiden hinnat vaihtelevat merkittävästi. Tiedostojen siirron kustannuksia vertaillaan epäsuorasti siirretyn internet- liikenteen avulla. Jokaiselle protokollalle laskettiin kuinka paljon ne kasvattavat tiedoston kokoa siirrettäessä. SFTP- prokolla kasvatti tiedoston kokoa kaikista eniten.

Tehokkuusvertailussa vertaillaan useaa eri asiaa tekemällä useita testejä. Tiedonsiirrossa huomattiin, että suojaamttomat protokollat ovat hieman nopeampia, mitä suojatut, mutta erot olivat erittäin pieniä. Tiedostojen talletuksessa huomattiin että Cosmos- tietokanta ei sovellu ollenkaan binääritiedostojen tallentamiseen. Talletusratkaisut, joita voitiin käyttää sisäverkossa, oli huomattavasti nopeampia kuin ulkoverkossa olevat talletusratkaisut.

Tietoturvavertailussa tutkittiin TLS-salausprotokollan tietoturvallisuutta tiedostonsiirrossa sekä vertailtiin tarvittavia palomuuriasetuksia eri siirtoprotokollilla. Azuren pääsynhallinnan periaatteet käytiin läpi tässä osiossa sekä Azuren talletusratkaisuissa ja pääsynhallinnassa havaitut tietoturvaongelmat. Lopuksi vielä vertailtiin pilvialustoista tehtyjä tietoturvatutkimuksia. Tutkimuksissa havaittiin että suurin syy tietoturvaongelmiin pilvialustoilla kehitetyissä sovelluksissa liittyy inhimillisiin virheisiin ja pilvialustojen tekniset tietoturvaongelmat ovat erittäin harvinaisia.

abstract

Teemu Pöytäniemi: File transfer and file storages on the Azure cloud platform

Master thesis

Tampere University

Master of engineering, information technology

9/2024

The thesis examines file transfer and file storage on the Azure cloud platform. File transfer and file storage are examined from three perspectives: cost, performance and security. File transfer is compared using different file transfer protocols. The thesis compares the following non-secure protocols: HTTP, WebDav, WebSocket and FTP. In addition, secure versions of all non-secure protocols are included, as well as the secure SFTP protocol. SQL database, NOSQL database, Object Storage, Block Storage and File Storage are used to store the files. The theoretical part of the information security section deals with RSA and TLS encryption.

For the thesis, a testing environment was set up in Azure. SFTP, FTPS and WebDav servers were set up on the Azure virtual machine for data transfer. A .NET server was set up for HTTP and WebSocket traffic. Angular front-end was used to test HTTP and WebSocket traffic from the browser. C# and Python libraries were used to test data transfer using different protocols. Postgres Database, Cosmos Database, Azure Blob, Azure Files and Azure Managed Disks were used to store the files.

The cost comparison compares the prices of Azure's storage solutions using three different storage sizes: 16GB, 245GB and 1 TB. The prices of the storage solutions vary significantly. File transfer costs are compared using indirectly method by measuring transferred internet traffic. For each protocol, it was calculated how much they increase the file size when transferring. The SFTP protocol increased the file size the most.

A performance comparison compares several different things by carrying out several tests. In file transfer, it was found that the unprotected protocols were slightly faster than the protected protocols, but the differences were very small. For file storage, it was found that the Cosmos database is not at all suitable for storing binary files. The storage solutions that could be used on the internal network were significantly faster than the storage solutions on the external network.

The security comparison examined the security of the TLS encryption protocol for file transfer and compared the required firewall settings for different transfer protocols. The principles of Azure's access management were discussed in this section, as well as the security issues identified in Azure's storage and access management solutions. Finally, security studies on cloud platforms were compared. The studies found that the main cause of security problems in applications developed on cloud platforms is related to human error and that technical security problems in cloud platforms are very rare.

ALKUSANAT

Diplomityö on tehty mielenkiinnosta pilvialustoja kohtaan ja se ei liity suoraan minkään yrityksen projektiin. Ajatus työnaiheesta tuli yhdestä terveydenhuollon projektista, jossa ladataan aivokuvia pilvialustalle.

Haluan kiittää Mika Välimäkeä neuvoista ja työn ohjauksesta.

Tampereella, 3.9.2024

Teemu Pöytäniemi

Sisällysluettelo

[1. Johdanto 1](#_Toc175367990)

[2. Teoria 3](#_Toc175367991)

[2.1 Tiedostojen siirto 3](#_Toc175367992)

[2.1.1 HTTP 3](#_Toc175367993)

[2.1.2 WebDav 6](#_Toc175367994)

[2.1.3 WebSocket 7](#_Toc175367995)

[2.1.4 FTP ja FTPS 11](#_Toc175367996)

[2.1.5 SFTP 13](#_Toc175367997)

[2.2 Tiedostojen talletus 15](#_Toc175367998)

[2.2.1 SQL-tietokannat 15](#_Toc175367999)

[2.2.2 NOSQL 16](#_Toc175368000)

[2.2.3 Object Storage 16](#_Toc175368001)

[2.2.4 Block storage 19](#_Toc175368002)

[2.2.5 File storage 20](#_Toc175368003)

[2.3 Pilvialustat 22](#_Toc175368004)

[2.3.1 Microsoft Azure 22](#_Toc175368005)

[2.4 Tietoturva 23](#_Toc175368006)

[2.4.1 RSA 23](#_Toc175368007)

[2.4.2 TLS 24](#_Toc175368008)

[3. Projektin esittely 30](#_Toc175368009)

[4. TULOKSET 31](#_Toc175368010)

[4.1 Tiedostojen siirto 32](#_Toc175368011)

[4.2 Tiedostojen talletus 34](#_Toc175368012)

[5. Projektien tulokset 35](#_Toc175368013)

[5.1 Kustannusvertailu 35](#_Toc175368014)

[5.1.1 Pilvialustojen talletusratkaisujen kustannukset 35](#_Toc175368015)

[5.1.2 Tiedostojen siirron kustannukset 41](#_Toc175368016)

[5.2 Tehokkuusvertailu 43](#_Toc175368017)

[5.3 Tietoturvavertailu 50](#_Toc175368018)

[6. yhteenveto 56](#_Toc175368019)

[Lähteet 58](#_Toc175368020)

LYHENTEET JA MERKINNÄT

ADD engl, Azure Active Directory, Microsoftin pääsynhallintajärjestelmä

AES engl, Advanced Encryption Standard, lohkosalaus

API engl, Application Programming Interface, ohjelman rajapinta

ASCII engl, American Standard Code for Information Interchange,

merkistökoodaus

AWS engl, Amazon Web Services, pilvialusta

C# ohjelmointikieli

DH Diffie–Hellman, salausalgoritmi

DNS engl, Domain Name System, palvelu IP-osoitteiden löytämiseen

EC2 Elastic Compute Cloud, AWS:n virtuaalikonepalvelu

FCP engl, Fibre Channel Protocol, verkkoprotokolla

FTP engl, File Transfer Protocol, tiedonsiirtoprotokolla

FTPS engl, File Transfer Protocol Secure, tiedonsiirtoprotokolla

HDD engl, Hard Disk Drive, talletusratkaisu

HTTP engl, Hypertext Transfer Protocol, tiedonsiirtoprotokolla¨

HTTPS engl, Hypertext Transfer Protocol Secure, tiedonsiirtoprotokolla

IDA engl, Information Dispersal Algoritmia, talletusalgoritmi

IO engl, Input Output, sisään ja ulos menevä liikenne

IOPS engl, Input Output Operations Per Second, IO- operaatioiden määrä sekunnissa

IP engl, Internet Protocol, tiedonsiirtoprotokolla

JSON engl, JavaScript Object Notation, dataformaatti

NIST engl, National Institute of Standards and Technology, Yhdysvaltain kansallinen standardisointi- ja teknologiainstituutti

NOSQL engl, Not only SQL, tietokantamalli

PCAP engl, Packet Capture, verkkokäytäntö

QUIC engl, general-purpose transport layer network protocol,

tiedonsiirtoprotokolla

RC4 engl, Rivest Cipher 4, virtasalaus

REST engl, Representational State Transfer, rajapinta-arkkitehtuuri

RSA Rivest–Shamir–Adleman, julkisen avaimen salausjärjestelmä

SaaS engl, software as a service, sovellus joka on tarjottu palveluna

SAN engl, Storage Area Network, tietoverkko

SCSI engl, Small Computer System Interface, kommunikointiprotokolla

SFTP engl, Secure File Transfer Protocol, tiedonsiirtoprotokolla

SPA engl, Single Page Application,selainkäyttöliittymä

SSD engl, Solid State Drive, talletusratkaisu

SSH engl, Secure Shell, tietoliikenneprotokolla

SSL engl, Secure Sockets Layer, salausprotokolla

SQL engl, Structured Query Language, kyselykieli

TCP engl, Transmission Control Protocol, tiedonsiirtoprotokolla

TLS engl, Transport Layer Security, salausprotokolla

UDP engl, User Datagram Protocol, tiedonsiirtoprotokolla

URL engl. Uniform Resource Locator, verkkosivun osoite

UTF-8 engl, Unicode Transformation Format-8, merkistökoodaus

WAL engl, Write Ahead Logging, tietokantojen protokolla

WebDav engl, Web-based Distributed Authoring and Versioning,

Tiedonsiirtoprotokolla

XML engl, Extensible Markup Language, dataformaatti

XSS engl, Cross-Site Scripting, hyökkäys uhrin selaimeen

.

# Johdanto

Moderni kaupallinen pilvilaskentainfraktuuri sai alkunsa vuonna 2002, kun Amazon Web Services (AWS) julkaistiin. Vuonna 2006 Amazon esitteli Simple Storage Service (S3) -palvelun ja Elastic Compute Cloud (EC2) -palvelun. Nämä palvelut olivat ensimmäisiä, jotka käyttivät palvelinvirtualisointia ja mahdollistivat asikkaiden vuokrata palvelimia heidän sovellustensa infrastruktuuriksi.[1]

Vaikka pilvipalveluiden kasvu alussa oli hidasta, viimeisten 15 vuoden aikana pilvipalvelut ovat laajentuneet merkittävästi. Google, Microsoft, Oracle ja IBM julkaisi omat pilvialustansa vuosien 2010 ja 2013 väilissä. Tämä auttoi saamaan pilvipalvelut yleisesti saataville ja pilvialustojen käyttö lisäänty merkittävästi. Nykyään pilvipalvelut ovat vallanneet suuren osan yrityksistä ja tuoreimmissa tutkimuksissa on havaittu, että yli 90% yrityksistä käyttää pilvipalveluita. [2][3]

Pilvilaskenta-alustat voidaan jakaa kolmeen ryhmään: julkiseen, yksityiseen ja hybridiin. Julkiset alustat ovat yleisin pilvipalveluiden alusta. Niissä pilviresurssit (kuten palvelimet ja tallennustila) ovat kolmannen osapuolen pilvipalveluntarjoajan omistamia ja ylläpitämiä, ja ne toimitetaan Internetin kautta. Microsoft Azure ja AWS ovat esimerkkejä julkisesta pilvestä. Julkisessa pilvessä samat laitteistot, tallennustilat ja verkkolaitteet jaetaan muiden organisaatioiden tai käyttäjien kanssa. Yksityinen pilvi koostuu pilvipalveluista, joita käyttää yksinomaan yksi yritys tai organisaatio. Yksityinen pilvi voi sijaita fyysisesti omassa palvelinkeskuksessa tai sitä voi isännöidä kolmannen osapuolen palveluntarjoaja. Hybridipilvi on alusta, joka yhdistää paikallisen infrastruktuurin tai yksityisen pilvialustan julkiseen pilvialustaan. Hybridialustat mahdollistavat tietojen ja sovellusten liikkumisen kahden ympäristön välillä. Monet organisaatiot valitsevat hybridipilvilähestymistavan liiketoiminnan tarpeiden vuoksi, kuten säännösten ja tietojen riippumattomuusvaatimusten täyttämisen, paikan päällä tapahtuvan teknologiainvestoinnin täyden hyödyn hyödyntämisen tai alhaisen viiveen ongelmien ratkaisemisen vuoksi. [4]

Tässä työssä tutkitaan pelkästään julkisen pilven alustoja, erityisesti Microsoftin Azurea. Työssä tutkitaan tiedostojen siirtämistä ja tallentamista Azure -pilvialustalla sekä vertaillaan julkisista pilvialustoista tehtyjä tietoturvatutkimuksia. Tutkimuksissa on arvioitu, että julkisten pilvialustojen talletusratkaisuiden liikevaihto vuonna 2023 olisi noin 99 biljoona Yhdysvaltain dollaria. Liikevaihdon ennuste vuodelle 2028 olisi 235 biljoonaa Yhdysvaltain dollaria. [5] Pilvialustojen talletusratkaisuiden suuri kysyntä ja arviot kysynnän kasvusta olivat yksi motivaatio työn tekemiseksi. Työ rajoittuu pelkästään binäärimuotoisiin tiedostoihin, eikä työssä käsitellä erikseen esimerkiksi tekstitiedostoja. Työn teoriaosuudessa tiedostojen siirrossa esitellään kaikki yleisesti käytetyt protokollat, joiden avulla tiedostot voidaan siirtää pilvialustalle. Tiedostojen talletuksen osuudessa esitellään pääarkkitehtuurit julkisten pilvialustojen tarjoamille talletusratkaisulle. Tietoturvaosuudessa käsitellään Rivest–Shamir–Adleman (RSA) -salausjärjestelmää sekä Transport Layer Security (TLS) -salausprotokollaa.

Työn tutkimusosassa käsitellään tiedostojen siirtoa ja talletusta kolmesta eri näkökulmasta: kustannukset, suorituskyky ja tietoturva. Tutkimusta varten Azure -pilvialustalle kehitettiin vaadittavat palvelut tutkittavien asioiden testausta varten. Azureen kehitettiin tiedonsiirtoprotokollille vaadittavat palvelimet, jotka integroitiin käytettäviin talletusratkaisuihin. Tiedostojen siirrossa asikasohjelmana käytetään selaimessa toimivaa käyttöliittymää sekä ohjelmointikielien kirjastoja. Tiedostojen siirtämisen sekä talletusten suorituskyvyn mittaaminen suoritetaan usealla erillaisella suorituskykytestillä. Talletusratkaisuiden kustannuksia vertaillaan kolmella eri talletuskoon avulla. Työssä pohditaan millaisia käyttötapauksia varten eri talletusratkaisut ovat kehitetty, jotta niiden käyttäminen olisi mahdollisimman edullista. Tiedostojen siirron kustannuksia vertaillaan epäsuorasti siirretyn internet -liikenteen avulla. Internet -liikenettä tutkitaan WireShark nimisen avoimen lähdekoodin pakettianalysaattorin avulla. Tietoturvaosuudessa esitellään suojaamattomien tiedonsiirtoprotokollien ongelmat ja tutkitaan TLS-salausprotokollan tietoturvallisuutta. TLS-salausprotokollaa käytetään suojaamaan salaamattomat tiedonsiirtoprotokollat. Tässä osuudessa esitetään Azuren pääsynhallinnan mekanismit sekä esitellään Azuren palveluiden tietoturvahaavoittuvuuksia. Lopuksi vertaillaan muutamaa tutkimusta, jotka käsittelevät julkisille pilvialustoille kehitettyjen sovellusten tietoturvasta.

# Teoria

## Tiedostojen siirto

Tässä luvussa käsitellään protokollia, joiden avulla tiedostoja voidaan siirtää pilvialustoille. Ensimmäisessä aliluvussa käsitellään selainten mahdollistamia lataustapoja, ja sen jälkeen siirrytään protokolliin, joita erilliset asiakasohjelmat tarvitsevat.

### HTTP

Hypertext Transfer Protocol (HTTP) on World Wide Webin (WWW) perusta, ja sitä käytetään verkkosivujen lataamiseen hypertekstilinkkien avulla. Hyperteksti on tekstiä, joka näytetään tietokoneen näytöllä tai muussa elektronisessa laitteessa, ja jossa on viittauksia hyperlinkeillä muihin teksteihin, joihin lukija pääsee välittömästi käsiksi. WWW käyttää Hypertext Markup Language (HTML) -tiedostoja määrittämään hypertekstitiedostojen rakenteen ja muotoilun. HTML on standardoitu merkintäkieli asiakirjoille, jotka on suunniteltu näytettäväksi verkkoselaimessa.[6]

HTTP on asiakas-palvelin-protokolla, jossa asiakasohjelma, tekee palvelimelle pyynnön, johon palvelin vastaa. HTTP:n yleisin asikasohjelma on verkkoselain, muita yleisiä asikasohjelmia ovat mobiili- ja työpöytäsovellukset. Asiakasohjelma lähettää HTTP-pyynnön palvelimelle, joka vastaa asiakasohjelman pyyntöön HTTP-vastauksella. HTTP- Vastaus sisältää pyydetyt tiedot tai resurssit, kuten verkkosivun, kuvan tai videon. HTTP on tilaton protokolla, mikä tarkoittaa, että jokainen pyyntö ja vastaus eivätkä ole riippuvaisia ​​aiemmista pyynnöistä tai vastauksista.[7]

HTTP- pyynnöt koostuvat seuraavista osista: metodi, polku, versio, otsikot sekä runko. Metodi määrittää toiminnon, minkä asikas haluaa suorittaa. Yleisimmät metodit ovat GET ja POST. GET -metodin avulla asiakas pystyy noutamaan tiedostoja palvelimelta ja POST -metodin avulla asikas pystyy lähettään dataa palvelimelle. Muut yleiset metodit ovat DELETE, PUT, PATCH, OPTIONS ja HEAD. Polku määrittää halutun palvelimen sijainnin verkossa. Polun määrittämiseen käytetään Uniform Resource Locator (URL) -osoitteita. Otsikot ovat avain-arvo -pareja tekstimuodossa, joiden avulla asikas ja palvelin voivat välittää lisätietoja pyyntöjen ja vastauksien mukana. Pyynnön runko on se osa, joka sisältää pyynnön siirtämät tiedot palvelimelle. Runko sisältää esimerkiksi kirjautumistiedot tai palvelimelle siirrettävän tiedoston. Versio on pyynnön käyttämä HTTP -protokollan versio.[8]

HTTP -vastaukset puolestaan pitävät sisällään tilakoodin, tilaviestin, version, otsikot ja rungon. Tilakoodi on kolminumeroinen numero, joka ilmoittaa, onko tietty HTTP -pyyntö saatu onnistuneesti päätökseen. Tilakoodit voidaan ryhmitellä viiteen ryhmään pyynnön onnistmisen persuteella: Informatiiviset vastaukset, onnistunnet vastaukset, uudelleenohjausviestit, asiakas virheet ja palvelin virheet. Tilaviesti on lyhyt tekstikuvaus tilakoodista. Versio, otsikot ja runko toimivat samalla tavalla kuin HHTP -pyynnössä.[8][9]

HTTP-protokollasta on julkaistu useita versioita: HTTP/0.9, HTTP/1.0, HTTP/1.1, HTTP/2 ja HTTP/3. HTTP -protokolla sai alkunsa 1980-luvun lopulla ja ensimmäinen internet palvelin kehitettiin CERNissä 1990. Vuonna 1991 ensimmäinen virallinen dokumentoitu HTTP-versio julkaistiin. Versio sai nimen HTTP/0.9 ja se tukee vain GET-pyyntöä, jonka avulla asiakkaat voivat noutaa HTML-asiakirjoja palvelimelta, mutta ei tue muita tiedostomuotoja. HTTP 1.0 julkaistiin vuonna 1996 ja se käyttää yksinkertaista pyyntö-vastausmallia, jossa asiakas lähettää pyynnön palvelimelle ja palvelin lähettää vastauksen takaisin asiakkaalle. Protokolla tukee GET, POST ja HEAD -pyyntöjä, ja protokollaan lisättiin otsikot sekä tilakoodit. Otsikoiden avulla protokollalla pystyy siirtämään muita tiedostotyyppejä kuin HTML -dokumentteja. Pyyntö- ja vastausviestit lähetetään verkon yli pelkkänä tekstinä, jossa otsikko ja teksti erotetaan tyhjällä rivillä. Jokaiselle pyynnölle luodaan oma TCP-yhteys ja vastauksen saapuessa TCP yhteys katkaistaan. HTTP/1.1 julkaistiin vuonna 1997 ja siitä julkaistiin päivitetty versio vuonna 1999. HTTP 1.1 tukee pysyviä yhteyksiä, jotka mahdollistavat useiden pyyntöjen ja vastausten lähettämisen peräkkäin saman TCP-yhteyden kautta. Tämä vähentää uusien yhteyksien luomista jokaista pyyntöä varten ja parantaa protokollan suorituskykyä. HTTP/1.1 yksi uusi ominaisuus on HTTP-putki, jonka avulla voidaan lähettää useita HTTP-pyyntöjä yhden TCP-yhteyden kautta odottamatta pyyntöjen vastauksia. Pyyntöjen ketjuttaminen johtaa huomattavaan parannukseen HTML-sivujen latausajoissa, erityisesti korkean viiveen yhteyksissä. Palvelimen on kuitenkin lähetettävä vastauksensa samassa järjestyksessä kuin pyynnöt vastaanotettiin. HTTP/1.1 suurin suorituskykyongelma on head-of-line (HOL)- esto . Se syntyy kun samaan kohteeseen on menossa useampi pyyntö, mutta jonon kärjessä oleva pyyntö, joka ei voi noutaa vaadittua resurssiaan, estää kaikki sen takana olevat pyynnöt. HOL- esto tapahtuu myös HTTP-putkessa, koska putkessa lähetettyihin pyyntöihin pitää vastata samassa järjestyksessä. Erillisten rinnakkaisten TCP-yhteyksien lisääminen voi helpottaa ongelmaa, mutta asiakkaan ja palvelimen välisten samanaikaisten TCP-yhteyksien määrä on rajoitettu, ja jokainen uusi yhteys vaatii huomattavia resursseja.[10][11][12]

HTTP/2 on päivitetty versio HTTP/1.1-protokollasta, joka otettiin käyttöön vuonna 2015. Se on suunniteltu korjaamaan joitakin HTTP/1.1:n rajoituksia ja suorituskykyongelmia, sekä parantamaan verkkoviestinnän nopeutta ja tehokkuutta. Yksi merkittävimmistä HTTP/1.1:n ja HTTP/2:n erottavista ominaisuuksista on binäärikehystyskerros. HTTP/1.1 pitää kaikki pyynnöt ja vastaukset pelkästään tekstimuodossa. HTTP/2 käyttää binaarikehystyskerrosta kaikkien viestien kapseloimiseen binäärimuotoon säilyttäen silti HTTP-semantiikan. Viestien muuntaminen binäärimuotoon sallii HTTP/2:n käyttää uusia siirtotapoja tiedon toimittamiseen, joita ei ole saatavilla HTTP/1.1:ssä. HTTP/2:ssa binäärinen kehystyskerros koodaa pyynnöt ja vastaukset ja leikkaa ne pienemmiksi tietopaketeiksi, mikä lisää huomattavasti tiedonsiirron nopeutta. Toisin kuin HTTP/1.1, jonka on käytettävä useita TCP-yhteyksiä HOL-eston vaikutuksen vähentämiseksi, HTTP/2 muodostaa yhden yhteyden kahden palvelimen välille. Tässä yhteydessä on useita tietovirtoja. Jokainen tietovirta koostuu useista viesteistä pyyntö/vastaus-muodossa. Jokainen näistä viesteistä jakaantuu vielä pienempiin yksiköihin, joita kutsutaan kehyksiksi. Yhteys koostuu joukosta binäärikoodattuja kehyksiä, joista jokainen on merkitty tiettyyn tietovirtaan. Tietovirran tunnisteet sallivat yhteyden lomittaa kehykset siirron aikana ja koota ne uudelleen toisessa päässä. Lomitetut pyynnöt ja vastaukset voivat toimia rinnakkain estämättä niiden takana olevia viestejä. Tätä prosessia kutsutaan multipleksaukseksi ja se ratkaisee HTTP/1.1:n HOL- eston. Yksi mahdollinen ongelma on, että useat tietovirrat ovat riippuvaisia tietyn tietovirran valmistumisesta ja joutuvat odottamaan sen valmistumista. Ratkaisuna ongelmaan on tietovirtojen priorisointi, jonka avulla tietovirralle voidaan määrittää prioriteetti ja riippuvuussuhde muihin tietovirtoihin. Palvelin käyttää näitä tietoja riippuvuuspuun luomiseen, jonka avulla palvelin voi määrittää järjestyksen, jossa tietovirtojen kehykset lähetetään.[10][11]

HTTP:n uusin versio HTTP/3 on Googlen kehittämä, ja se julkaistiin vuonna 2015. Suurin muutos versioon HTTP/2 nähden on taustalla olevan siirtokerroksen verkkoprotokollan TCP:n vaihtaminen Quick UDP Internet Connections (QUIC) -protokollaan. QUIC on Googlen vuonna 2012 julkaisema protokolla, joka käyttää siirtokerroksessa UDP:ta TCP:n sijaan. QUIC on multipleksoitu viestintäalgoritmi, joka kehitettiin paremmaksi algoritmiksi kuin HTTP/2:n käyttämä multipleksi. Vaikka HTTP/2 poistaa HOL- eston HTTP-protokollasta, niin TCP-protokollalla on oma HOL- esto. Jos TCP hukkaa paketin yhdessä HTTP/2 luomassa tietovirrassa, kaikkien muiden tietovirtojen kehysten siirtäminen samassa TCP-yhteydessä estyy kunnes hukatut paketit ovat saatu uudelleen lähetettyä. QUIC-tietovirran data jaetaan pienempiin osiin, joita kutsutaan kehyksiksi. Kehykset lähetetään UDP-paketteina ja ne sisältävät tietovirran tunnuksen ja järjestysnumeron. Vastaanottaja voi järjestää ne uudelleen, jos ne vastaanotetaan eri järjestyksessä mitä lähetettiin. Jos paketti katoaa, QUIC lähettää sen uudelleen aikakatkaisun jälkeen. Logiikka on sama mitä TCP tekee HTTP/2 -protokolassa, mutta se ei vaikuta yhteyden muihin tietovirtoihin.[13][14]

### WebDav

Web Distributed Authoring and Versioning (WebDav) on HTTP/1.1:n laajennus. WebDAV-protokolla esiteltiin alun perin vuonna 1996, ja se standardointiin ensimmäistä kertaa vuonna 1999. Protokollan nykyinen versio on vuodelta 2007. WebDav mahdollistaa HTTP- palvelimen toimimisen tiedostopalvelimena mahdollistamalla enemmän kuin yhden asiakkaan työskentelemisen saman tiedoston parissa samaan aikaan. WebDAV laajentaa HTTP-otsikoiden ja -metodien standardijoukkoa. WebDav:n avulla voi luoda, siirtää, muokata poistaa ja kopioida tiedostoja ja kansioita. WebDAV- palvelimet mahdollistavat tiedostojen versioiden seurannan ja palvelimet on jaettu kahteen luokkaan versioiden seurannan perusteella: luokka 1 ja luokka 2. Luokan 1 WebDAV-palvelimet tarjoavat perushallintaominaisuuksia, kuten mahdollisuuden luoda, kopioida, siirtää tai poistaa tiedostoja ja kansioita. Monet asiakasohjelmat pitävät luokan 1 WebDAV -palvelimia vain luku -muotoisina, koska ne eivät voi suojata tiedostoja samanaikaisilta muutoksilta. Luokan 2 WebDAV-palvelimet voivat lukita tiedostoja ja sallivat tiedostojen samanaikaisen muokkauksen.[15][16][17]

WebDav lisää seuraavat HTTP -metodit HTTP/1.1 -standardiin: PROPFIND, PROPPATCH, COPY, MOVE, MKCOL, LOCK ja UNLOCK. WebDav käyttää PUT- ja DELETE -metodeja, jotka löytyvät jo valmiiksi HTTP -standardista. Protokolla käyttää tiedostoista ja kansioista resurssi nimitystä. PROPFIND -metodilla pystyy hakemaan tietoja resursseista ja PROPPATCH -metodi mahdollistaa niiden muokkaamisen. LOCK -metodi lukitsee halutun resurssin, sillain että toiset käyttäjät eivät pysty tekemään muutoksia siihen. UNLOCK -metodi purkaa lukituksen. COPY -metodi kopioi resurssin palvelimella ja MOVE -metodi siirtää sen. resurssin pystyy lataamaan palvilimelle PUT metodilla ja poistamaan DELETE -metodilla ja lataamaan GET -metodilla. [17]

Kuva 1 esittää esimerkkitapausta asiakkaan ja palvelimen välisestä liikenteessä, jossa asiakas lukitsee ensin tiedoston, ja muokkaa sitä ja sen jälkeen, ja purkaa lukituksen lopuksi.

A screenshot of a computer

Description automatically generated

Kuva 1 Esimerkki WebDav -latauksesta

### WebSocket

WebSocket -protokolla luo kaksisuuntaisen realiaikaisen yhteyden asiakkaan ja palvelimen välille. Tämä mahdollistaa sen, että palvelin voi lähettää dataa asiakkaalle ilman asiakkaan erillistä pyyntöä. WebSocket on tilallinen protokolla, mikä tarkoittaa, että asiakkaan ja palvelimen välinen yhteys pysyy elossa, kunnes jompikumpi osapuoli lopettaa sen. Kun asiakas tai palvelin on sulkenut yhteyden, yhteys katkeaa molemmista päistä. WebSockets toimii TCP:n yli toimivana siirtokerroksena. Websocket-protokolla muistuttaa TCP-protokollaa, mutta protokollat eroavat kuitenkin toisistaan ja WebSocketilla ei voi muodostaa yhteyttä TCP-serveriin. WebSocketin pääasiallinen asiakasohjelma on selain, mutta muillakin asikasohjelmilla voidaan luoda yhteys palvelimeen. WebSocket API on selaimen API, joka tarjoaa mekanismin yhteyden muodostamiseen selaimesta WebSocket -palvelimeen JavaScript ohjelmointikielen avulla.[18][19]

Kuva 2 esittää WebSocketin arkkitehtuuri korkealla tasolla. Aluksi kättelyvaiheeseen käytetään HTTP – protokollaa, jonka jälkeen siirrytään WebSocket -protokollaan.



Kuva 2 webSocket-protokolla

Asiakas aloittaa kättelyprotokollan lähettämällä HTTP GET -pyyynnön palvelimelle. Pyynnössä pitää olla mukana tietyt HTTP-otsikot, jotta palvelin hyväksyy pyynnön. Upgrade- ja Connection -otsikon arvoilla kerrotaan palvelimelle, että yhteys halutaan muuttaa WebSocket -protokollaan. WebSocketin version pitää olla 13, joka on nykyinen versio WebSocket -standardista. Pyynnössä on myös mukana Sec-WebSocket-Key otsikko, jonka arvo on satunnainen 16-tavuinen merkkijono, joka on base-64 koodattu. Tämän merkkijonon generointi pitäisi olla kryptografisesti satunnainen ja sen arvoja ei pitäisi pystyä arvaamaan. Jos palvelin hyväksyy kättelyn, niin pyynnön HTTP-vastauskoodi on 101.[19]

Kun kättely on onnistuneesti tehty, siirrytään WebSocket-protokollaan. Kuva 3 esittää protokollassa liikkuvia kehyksiä, joiden avulla data siirretään. Kehyksessä olevilla lohkoilla on seuraavat merkitykset alkaen ensimmäisestä bitistä:

* FIN, 1 bitti: merkkaa onko kehys lähetettävän datan viimeien vai ei.
* RSV1, RSV2 ja RSV3 1 bitti kukin. Kaikki ovat nollia jos ei ole lisäosaa kommunikoinnin alussa sovittu.
* Opicode, 4 bittiä. Koodi joka merkkaa kehyksen tyypin
* MASK, 1 bitti. Merkkaa onko siirrettävä data naamioitu vai ei. Jos asiakas lähettää kehyksen tämä bitti täytyy olla 1
* Siirrettävän datan pituus, 7 bittiä.
* naamiointiavain, 32 bittiä jos Mask bitti on 1. Tällä avaimella asiakkaan on naamioitava data
* Siirrettävä data, jonka pituus on ilmoitettu siirrettävän datan pituus-arvolla. Se pitää sisällään lisäosa-datan sekä sovellus-datan. Lisäosa-datan pituus on 0 jos lisäosaa ei ole käytössä, muuten lisäosa määrittää tämän datakentän pituuden. Sovellus-datan pituus on siirrettävän datan pituus miinus lisäosa-datan pituus.



Kuva 3: WebSocket kehys

Kun asiakas lähettää kehyksen palvelimelle, MASK bitillä täytyy olla arvo 1, ja 32-bittinen naamiointiavain määriteltynä. Avaimen pitää olla jokaiselle kehykselle eri ja sen generoivan algoritmin kryptografisesti vahva, jotta tulevia arvoa ei voida arvata vanhojen arvojen pohjalta. Datan naamioiminen tapahtuu 8 bitin lohkoissa ja siinä käytetään jakojäännös- ja XOR- operaatioita.[19]

Datan naamioimisen tarkoitus on tehdä WebSocket-liikenteestä HTTP-liikenteestä poikkeavaa ja arvaamatonta. Muuten verkkoinfrastruktuurilaitteet, joita ei ole päivitetty ymmärtämään WebSocket-liikennettä, voivat luulla sitä virheellisesti normaaliksi HTTP-liikenteeksi ja avata tietoturva-aukkoja. Naamioimisella pystytään suojaamaan verkkoinfrastruktuurilaitteet, ja vaikka hyökkääjä pystyisi hallitsemaan asiakaspuolen sekä palvelimen koodia. Hyökkääjä ei pysty kuitenkaan hallitsemaan liikutettavan kehyksen arvoja, koska ne on naamioimisen ansiosta satunnaisia.[20]

4-bittisellä Opcode-koodilla pystytään merkkaamaan kehyksen tyyppi. Kehys voi olla datan siirtoon tarkoitettu kehys tai hallintakehys. Jos kehys on hallintakehys, sillä on sovittu hallintakoodi Opicode lohkossa ja mahdollisesti siirrettävää dataa. Mahdollisia komentoja on close, ping ja pong. Close -komento aloittaa yhteyden sulkemisprosessin, ja se voi sisältää datakentässä viestin, miksi yhteys halutaan sulkea. Kun Close-komento on vastaanotettu, siihen vastataan lähettämällä kehys, jonka komento on myös Close. Close-vastausta voi viivästyttää ja jatkaa datan lähettämistä, mutta silloin ei ole taetta että Close komennon lähettänyt osapuoli pitää yhteyden auki. Ping- komennon tarkoitus on tarkistaa että yhteyden toinen pää ottaa vielä vastaan liikennettä. Kun Ping -komento on vastaanotettu, siihen vastataan lähettämällä Pong -komento. Jos kehys on datan siirtokehys Opcode-koodilla hallitaan siirrettävän datan formaattia. Mahdollisia vaihtoehtoja ovat binääri ja teksti. Tekstin käyttämä merkistö on UTF-8. Opcoden avulla voidaan pilkkoa lähetettävä data useampaan kehykseen, jolloin kaikkien keysten datan formaatti pitää olla sama kuin ensimmäinen kehyksen. Ensimmäinen kehys määrittää Opcodella siirrettävän datan tyypin ja seuraavien kehysten Opcode määrittää että se on jatkoa edelliselle kehykselle. Datan viimeisen kehyksen FIN-bitin arvo on 1, joka kertoo että kyseessä on datan viimeinen kehys. Opicode-koodissa on vapaita komentoja lisäosia varten, joita ei ole määritelty WebSocket-protokollaan. Asiakas voi pyytää lisäosan käyttöä kättelyvaiheessa lisäämällä Sec-WebSocket-Extensions otsikon pyyntöön. Jos serveri hyväksyy lisäosia käyttöön, se lisää saman Sec-WebSocket-Extensions otsikon vastaukseen, johon se listaa hyväksytyt lisäosat.[19][20]

### FTP ja FTPS

FTP eli File Transfer Protocol on tiedostojen siirtoon käytetty protokolla. Protokolla toimii TCP/IP-yhteyden päällä ja se koostuu asiakkaasta ja palvelimesta. Kuvassa 2 on esitetty FTP protokollan arkkitehtuuri korkealla tasolla. Sekä asiakas että palvelin tarvitsevat ohjelmat, jotka toteuttavat protokollan. Asiakkaan ohjelma voi olla graafinen käyttöliittymä tai se voi toimia komentoriviltä.[21]



Kuva 4 FTP- protokolla

Protokolla käyttää kahta rinnakkaista TCP -yhteyttä: hallintayhteyttä ja datayhteyttä. Hallintayhteydellä lähetetään kaikki komennot liittyen käyttäjien hallintaan ja tiedonsiirtoon, ja se on totetutettu Telnet-protokollalla. Telnet on vanha sanomamuotoinen protokolla, joka toimii TCP-prokollan päällä. Telnetillä pystyy muodostamaan yhteyden etätietokoneelle ja sitä on käytetty aikoinaan mm. sähköpostien lähetykseen. Telnet ei salaa sanomiaan, minkä vuoksi protokolla on tietoturvaltaan heikko. Telnetin käytöstä on luovuttu moderneissa järjestelmissä ja se on korvattu turvallisimmilla protokollilla, kuten SSH:lla. Hallintayhteydellä määritellään datayhteyden ominaisuudet ja sillä avataan haluttu datayhteys.[21]

Datayhteydellä siirretään tiedostot. Yhdellä aktiivisella yhteydellä voidaan siirtää vain yhtä tiedostoa samanaikaisesti. Datayhteys siirtää datan aina 8-bittisinä tavuina, joita kutsutaan siirtotavun kooksi. FTP tukee muutamaa eri tiedonsiirtotyyppiä, joina se pystyy esittämään siirrettevän datan. Oletustyyppi on ASCII, jota kaikkien FTP-toteusten pitää tukea. ASCII on tekstin merkistö, jossa jokainen merkki esitetään 7 bittisinä numerona. Tiedosto lähetetään 8-bittisenä NVT ASCII:na jossa eniten merkitsevä bitti on nolla ja sen perässä 7-bittinen ASCII koodi. Vaihtoehtoinen tyyppi lähettää tekstitiedostoja on EBCDIC. Se on 8-bittinen merkistö, joka on pääasiassa käytössä IBM:n suurtietokoneissa. ASCII ja EBCDIC ottavat myös toisen vapaaehtoisen parametrin joka määrittää tekstin pystysuuntaista asettelua ja sitä käytetään tekstin tulostuksen muotoiluun. Oletusarvo Non-print ei ota kantaa muotoiluun. Telnet format control -parametriksi annettuna olettaa että tekstistä löytyy vertikaalisia tekstin hallinta komentoja kuten LF ja NL. Carriage control olettaa että tekstistä löytyy ASA (FORTRAN) vertikaali hallinta merkkejä ja hallitsee niiden avulla tekstin muotoilua. Image-tyyppi on binääritiedostojen lähettämiseen tarkoitettu tyyppi. Tiedosto luetaan biteiksi ja bitit lätetään 8-bittisinä tavuina. Lähetyksen loppuun saatetaan tarvita täytetttä jos tiedoston bitit eivät ole kahdeksalla jaollisia. Täyte on pelkästään nollia, ja asiakkaan ja palvelimen täytyy sopia kuinka täyte tunnistetaan oikeasta datasta. Local-tyyppi on image-tyypin laajennos ja sen avulla voidaan muuttaa lähetettävän datan loogisen tavun kokoa annettavalla parametrillä.[21]

FTP tukee kolmea eri tiedostonvälitysrakennetta. Tiedostorakenne, tallennerakenne ja sivurakenne: Tiedostorakenteessa tiedostolla ei oleteta olevan mitään sisäistä rakennetta, vaan tiedosto lähetetään binäärimuodossa. Tallennerakenteessa tiedoston oletetaan koostuvan toistuvista pienemmistä osista. Sivurakenteessa tiedoston oletetaan koostuvan sivuista, jotka ovat indeksoitu. Tätä rakennetta käytetään epäjatkuvien tiedostotyyppien kanssa jotka voivat sisältää useampia sisäisiä tiedostoja jotka tunnistetaan tiedostossa olevalla otsikkodatalla.[21]

Datan siirtämisen aloittaminen edellyttää datayhteyden muodostamista, sekä siirtoparametrien sopimista. Aluksi asiakkaan pitää avata datayhteyden portti kuuntelemista varten ja lähettää pyyntösanoma hallintayhteydellä. Tämä sanoma pitää sisällään tiedon kumpaan suuntaa dataa ollaan siirtämässä. Palvelin vastaa pyyntöön ottamalla yhteyden asiakkaan datayhteyden porttiin. Seuraavaksi tiedoston siirtotapa pitää sopia. FTP tukee kolmea tapaa: stream, block ja Compressed. Stream lähettää tiedostot jatkuvana virtana. Jos tiedoston tiedostonvälitysrakenne on tiedostorakenne, niin tiedoston päättyminen merkataan sulkemalla datayhteys. Tallennerakenteessa käytetään kaksitavuista hallintakoodia merkkaamaan osien päättymistä tai tiedoston päättymistä. Block-siirrossa data siirretään lohkoina. Kullakin lohkolla on 3-tavuinen otsikko, jossa on tieto lohkon pituudesta sekä lohkon tyypistä. Lohkon tyypillä pystytään merkkaamaan lohkoon tieto siitä, että onko lohkotiedoston viimeinen, tallenteen viimeinen vai uudelleenkäynnistyksen merkki. Compressed-siirto käyttää yksinkertaista pakkausalgoritmia nimeltään run-length encoding joka lähettää datan lohkoissa kuten block-siirrossa. Pakkaamalla data siirrossa saadaan pienennettyä siirrettävän datan määrää, mutta yleensä pakkaus tehdään tehokkaimmilla algoritmeilla jo tiedostoa luodessa tai verkkotasolla modeemeiden toimesta. FTP-protokolla ei sisällä tapaa tarkistaa siirrettävän datan eheyttä, vaan se luottaa TCP-protokollaan. FTP sisältää kuitenkin protokollan tiedoston latauksen jatkamiseen, jos lataus on keskeytynyt. Latauksen jatkamista voi käyttää vain jos siirtotapa on block tai compressed. Latauksen jatkaminen merkataan lohkon otsikkoon sille varatulla koodilla paikassa, josta tiedoston lataus jatkuu.[21]

FTP-protokolla ei salaa kumpaakaan käyttämäänsä yhteyttä joten sitä pidetään nykyään tietoturvattomana. FTPS (FTP Secure) on FTP-protokollan tietoturvallisempi lisäosa, joka käyttää SSL- tai TLS-suojausprotokollaa. Suojauksen päälle kytkeminen voidaan toteuttaa kahdella eri tavalla: implisittisesti ja eksplisiittisesti. Implisiittinen tapa vaatii suojausta jo heti yhteyden luomisvaiheessa. Tämä rikkoo pelkästään FTP-protokollalla toimivien asiakasohjelmien ja palvelimien toimivuuden implisiittisesti suojausta vaativan vastinparinsa kanssa. Eksplitisiittisessä suojauksessa asiakkaan täytyy erikseen pyytää suojattua yhteyttä. Näin palvelin mahdollistaa kommunikoinnin sekä FTP- että FTPS -asiakkaiden kanssa. Eksplisitiittisesti määriteltäessä voidaan myös säätää mitkä yhteydet suojataan, kun implisiittisesti määriteltynä molemmat kanavat ovat suojattuna. Esimerkiksi jos siirrettävä data on jo valmiiksi salattu, pelkästään hallintayhteys voidaan salata.[21][22]

### SFTP

SFTP-protokolla on tiedostonsiirto- ja tiedostonhallintaprokolla, joka koostuu palvelimesta ja asiakasohjelmasta joka ottaa yhteyden palvelimeen. SFTP-prokolla toimii suojatun turvallisen kanavan avulla ja se suunniteltiin SSH 2-protokollan laajennokseksi. SFTP prokolla ei itsessään suojaa liikennettä, vaan se luottaa alla olevan kanavan suojaukseen. SFTP-yhteyden avulla voidaan siirtää tiedostoja ja tehdä tiedostonhallintaa. Vaikka SFTP-protokolla suunniteltiin SSH 2-prokollan laajennukseksi, sitä voidaan käyttää myös muiden protokollien avulla esimerkiksi SSH-1 protokollan avulla. SSH 1-prokollan käyttö aiheuttaa rajoitteita asiakasohjelmalle, koska sen pitää tietää SFTP-palvelimen binääritiedostojen sijainti palvelimella. SFTP-protokollasta on julkaistu useita versioita, joista käytetyin versio nykyään on versio 3 joka julkastiin vuonna 2001. Vuoden 2001 jälkeen protokollasta on julkaistu versiot 4, 5 ja 6, mutta ne eivät ole laajalti tuettuja ja useat SFTP-palvelimien kehittäjät ovat ilmoittaneet, että heillä ei ole aikomustakaan tukea uusimpia versioita tulevaisuudessa.[23][24]

SFPT-protokolla lähettää paketit verkon yli muodossa, jossa on alussa tavun kokoinen tila komennolle ja sen jälkeen tulee siirrettävä data. Kuvassa 5 on esitetty SFTP-protokollan arkkitehtuuri, ja siinä näkyy pakettien komentojen nimet. Asikas lähettää aluksi SSH\_FXP\_INIT-komennon jonka datana on haluttu versio protokollasta. Palvelin vastaa SSH\_FXP\_VERSION-komennolla, joka pitää sisällään palvelimen tukemat versiot. Palvelimen ja asiakasohjelman pitää tukea samaa versiota protokollasta, jotta yhteys pystytään muodostamaa. Tiedostojen käsitteleminen tapahtuu niille tarkoitetuilla komennoilla. Esimerkiksi SSH\_FXP\_OPEN komento avaa tiedoston tai luo uuden tyhjän tiedoston. Palvelin palauttaa SSH\_FXP\_OPEN kyselyyn SSH\_FXP\_HANDLE komennon jos SSH\_FXP\_OPEN kysely oli onnistunut. SSH\_FXP\_HANDLE palauttaa datana ID:n jonka avulla tiedosto voidaan siirtää asiakasohjelmalle SSH\_FXF\_READ lukukomennon avulla. Esimerkkejä muista komennoista : SSH\_FXF\_WRITE mahdollistaa tiedostojen kirjoittamisen palvelimelle, SSH\_FXP\_REMOVE mahdollistaa tiedostojen poistamisen palvelimelta.[24]Diagram

Description automatically generated

Kuva 5 SFTP-protokolla

## Tiedostojen talletus

Tässä luvussa käsitellään eri talletusrartkaisuita, joita pilvialustat yleisesti tarjoavat.

### SQL-tietokannat

SQL-tietokannat ovat relaatiotietokantoja, joissa tieto on tallennettuna tauluina, jotka koostuvat riveistä ja sarakkeista. Taulut perustuvat relattiomalliin, jonka määrittää predikaattilogikka. Taulu esittää yhden asian tiedot, sarakkeet kuvaavat kokonaisuuden arvoja ja rivit asian ilmentymiä. Tauluissa jokaisella rivillä on uniikki ID, joka pystyy muodostumaan joko yhdestä tai useammasta sarakkeen arvosta. Taulujen välille pystytään luomaan yhteyksia toisiinsa taulujen avainten avulla.[25]

Tietokannan käytön mahdollistavat tietokannan ohjausjärjestelmät. Ne mahdollistavat käyttäjän pääsyn tietokantaan, sekä sen päivittämisen ja muokkauksen. Tietokannan ohjausjärjestelmiä on maksullia ja avoimen lähdekoodin ilmaisia järjestelmiä. Suosituimmilla avoimen lähdekoodin ohjausjärjestilmillä on hyvin samallainen pääarkkitehtuuri, mutta toteutuksissa on pieniä eroavaisuuksia.Kuva 6 esittää Postgres tietokannan arkkitehtuuria. Postgres on yksi suosituimmista avoimen lähdekoodin tietokannosta, ja se on julkaistu vuonna 1995. Postgres toimii asikas-palvelin -mallilla, jossa asikasohjelma ottaa yhteyden ohjausjärjestelmään. Postgres luo jokaisella asikasyhteydelle oman palvelinprosessin. Kaikilla palveluprosesseilla on pääsy jaettuun muistiin, joka on RAM -muistissa. Tietokantakyselyt tehhdään jaettuun muistiin, joka on kopio levyllä olevasta tiedosta. Jaettumuisti on huomattavasti nopeampaa kuin levyltä luku. Lukuoperaatiot kohdisttuvat pelkästään jaettuun muistiin, mutta kirjoitus operaatiot tehdään Write ahead log (WAL) -logiin. Ensiksi muutokset tehdään jaettuun muistiin, jonka jälkeen WAL-kirjoittaja kirjoittaa muutokset levylle säännöllisenä taustaprosessina. Checkpointer- ja writer -prosessit päivittävät muutokset jaetusta muistista levylle sen jälkeen kun muutokset ovat tehty WAL- tiedostoon. Levylle kirjoitus tehdään taustaprosessina pienissä paloissa koska I/O -operaatio on raskas. Jaetusta muistissa pidetään tallessa poistettuja rivejä sekä muuttuneiden rivien vanhoja versioita, koska toiset palveluprosessit saattavat tarvita vanhoja arvoja. Autovacuum – prosessi poistaa vanhentuneet rivit jaetusta muistista.[26][27] Diagram

Description automatically generated

Kuva 6 Postgresin arkkitehtuuri

### NOSQL

NoSQL tarjoaa mahdollisuuden tallentaa datan tietokantaa muuna rakenteena kuin relaatiomallin tauluina. Ensimmäiset NoSQl tietokannat kehitettiin jo 1960 -luvulla, mutta nimitys NoSQL syntyi 2000 -luvun alussa. NoSQL kannoille ei ole yhtä määritelmää, vaan ne jakautuvat neljään pääkategoriaan: asiakirjatietokannat, avain-arvo -tietokannat, sarakepohjaiset tietokannat ja graafitietokannat. Asiakirjatietokannat tallentavat datan dokumentteina. Yleisimmin käytetyt tyypit dokumenteille ovat JSON, BSON ja XML. Tietokannan dokumentit voivat olla sisäkkäisiä ja tietyt elementit dokumentin sisällä voidaan indeksoida nopeamman haun mahdollistamiseksi. Asiakirjatietokannat käyttävät kokoelmia, jonne tallennetaan yleensä asiakirjoja, joilla on samankaltainen sisältö Asiakirjatietokannat tallentavat datan kokoelmiin, samaan tapaan kuin SQL -tietokannat tallentavat datan tauluihin. Asiakirjatietokannoissa on joustavat skeemat, mikä tarkoittaa, että kaikissa kokoelman asiakirjoissa ei tarvitse olla samoja kenttiä. Skeemavalidoinnin avulla kentille voidaan luoda validointisääntöjä, kuten sallitut tietotyypit ja arvoalueet ja pakottaa asiakirjojen kentät samoiksi. Dokumenttitietokannoissa on yleensä API tai kyselykieli, jonka avulla voidaan suorittaa operaatioita tietokantaan. Avain-arvo -tietokannat ovat tietokantatyyppejä, joissa data tallennetaan avain-arvo-muodossa ja optimoidaan tietojen lukemista ja kirjoittamista varten. Tietoja haetaan yksilöllisen avaimen tai useiden yksilöllisten avainten avulla. Arvot voivat sisältää minkä tahnsa tietotyypin. Yksi yleinen käyttötapaus Avain-arvo -tietokannoille on välimuistiin tallentaminen. Sarakepohjaiset tietokannat tallentavat datan tauluihin samaan tapaan kuin SQL -tietokannat. Ero syntyy näiden tietokantojen välille tavasta, mitenkä tietokannat tallentavat taulussa olevan datan levylle. Kuva 7 esittää SQL -tietokantojen ja Sarakepohjaisten tietokantojen tapaa tallentaa tauluissa oleva data levylle.

A screenshot of a computer

Description automatically generated

Kuva 7 SQL-tietokantojen ja sarakepohjaisten tietokantojen taulujen tallentamisen erot

SQL- tietokannat tallentavat yhden rivin tiedot muistiin samaan paikkaan ja ne optimoitu lukemaan ja kirjoittamaan rivejä nopeasti. Sarakepohjaiset tietokannat järjestävät tiedot sarakkeittain ja pitävät kaikki sarakkeeseen liittyvät tiedot muistissa vierekkäin ja ne on optimoitu lukemaan ja kirjoittamaan sarakkeita tehokkaasti. Sarakepohjaiset tietokannat ovat erityisen hyödyllisiä analyyttisissä kyselyissä, joissa skannataan tai aggregoidaan usein suuria tietokokonaisuuksia, mutta tarvitaan vain muutamia sarakkeita. Koska vain tarvittavat sarakkeet luetaan tallennustilasta, I/O-operaatioihin kuluva aika on paljon pienempi, mitä SQL -tietokannoissa. Vaikka sarakepohjaiset tietokannat soveltuvat erinomaisesti analytiikkaan, niiden tapa kirjoittaa tietoja vaikeuttaa niiden käyttämistä. Tietokantaan uuden rivin lisääminen vaatii useita kirjoitusoperaatioita, koska jokaisen sarakkeen arvo pitää kirjoittaa levylle eri paikkaan.

### Object Storage

Object Storage tarjoaa tallennusmahdollisuuden rakenteettomalle datalle. Tiedostojen tyypillä ja koolla ei ole merkitystä vaan kaikki tiedostot tallennetaan yhteen ämpäriin. Ämpärissä tiedostoilla ei ole minkään sortin kansiorakennetta, vaan jokaisella tiedostolla on ämpärissä uniikki id. Ämpäriin pääsee käsiksi HTTP REST-rajapinnan kautta, joka mahdollistaa tiedostojen hallinnan. Jokainen Object Storagen toteuttava taho voi luoda oman REST-rajapinnan, mutta käytännössä kaikki pilvialusta toteuttavat Amazon AWS S3 ämpärin rajapinnan. S3 oli ensimmäinen Object Storage pilvialustoilla ja sen rajapinta oli hyvin dokumentoitu, jonka seurauksena siitä tuli erittäin suosittu ja muutkin pilvialustat alkoivat tarjoamaan samaa rajapintaa. Object Storage tarjoaa erittäin hyvän skaalautuvuuden, koska ämpärin kokoa voidaan kasvattaa erittäin suureksi, jopa exatavujen suuruiseksi. Object Storage on myös erittäin kustannustehokas ratkaisu, kun vain käytetystä talletustilasta laskutetaan eikä yhtään ylimääräisestä varalla olevasta talletustilasta kuten normaaleissa levytalletusratkaisuissa yleensä joutuu maksamaan. Nykyään jokaisella suurella pilvialustalla on oma versio Object Storagesta ja niiden sisäiset toteutukset voivat vaihdella. Seuraavaksi on esitelty IBM:n toteutus asialle ja Objectect Storagen arkkitehtuuri on esitetty kuvassa 5. Manager tarjoaa toiminnot admin-käyttäjille ja sen avulla voidaan valvoa ja muokata järjestelmää. Accesser tarjoaa REST-rajapinnan käyttäjille, sekä hoitaa datan salauksen kirjoittaessa ja salauksen purun lukuvaiheessa. Slicestor on vastuussa tiedostosiivujen tallentamisesta, jotka se saa Accesserilta. Slicestoreiden ryhmää kutsutaan Device Setiksi ja Device Settien ryhmää puolestaan Storage pooliksi. Storage poolit voivat koostua Device Seteista, jotka sijaitsevat eri palvelinkeskuksissa. Ämpärit kuvataan sisäisessä arkkitehtuurissa Voulteina ja ne sijaitsevat Storage poolissa. Yksi Storage pool voi pitää sisällään useita Voulteja ja jokainen Voult pitää sisällään Slicestoreita jokaisesta Device Setistä

. 

Kuva 8 IBM Object Storage

Datan tallentamiseen käytetään Information Dispersal Algoritmia (IDA). Algoritmin perusidea on seuraava: Tallennettava datasta luodaan N kappaletta paloja. Tätä lukua kutsutaan algoritmin leveydeksi. Kuitenkaan kaikkia näitä paloja ei tarvita, jotta tiedosto pystytään lukemaan tai kirjoittamaan kokonaisuudessan, vaan pienempi määrä paloja riittää. Read Treshold kertoo määrän kuinka monta palasta tarvitaan alkuperäisen datan lukemiseksi ja Write Threshold kuinka monta palasta tarvitsee olla kirjoitettuna, jotta kirjoitusoperaatio merkataan suoritetuksi. Write Threshold on aina suurempi mitä Read Treshold.

Kuvassa 6 on esitetty tiedoston kirjoitusprosessi. Prosessi lähtee likkeelle Accesser Nodeista, jotka pilkkovat yli 4 MiB kokoiset tiedostot neljäksi erilliseksi siivuksi. Siivut viedään yksitellen loppuprosessin läpi tai pieni tiedosto viedään kerralla kokonaisena. Seuraavaksi dataan lisätään eheyden tarkistusta varten tarkistusarvo tiivistefunktiolla ja data voidaan salata lohkosaluksen avulla. Data viipaloidaan ensiksi Read Treshold arvon mukaisesti viipaleista ja näistä viipaleista luodaan algoritmin leveyden verran paloja joista tarvitsee kirjoittaa algoritmin Write Threshold- määrän mukaisesti paloja, jotta kirjoitusoperaatio merkataan suoritetuksi. Nämä palat lähetetään Slicestor Nodeille tallennettavaksi. Kun kirjoitusoperaatio on merkattu onnistuneesti suoritetuksi, puuttuvat palaset kirjoitetaan asynkroonisesti taustaprosessina. Lukuprosessi on päinvastainen suhteessa kirjoitusprosessiin. Slicestor Nodeilta luetaan Read Treshold- määrän mukainen määrä siivuja ja näistä siivuista palautetaan alkuperäinen tiedosto.



Kuva 9 Tiedoston kirjoitusprosessi

### Block storage

Block storage tallentaa datan vakiokokoisina paloina, jotka se tallentaa yksilöllisen tunnisteen avulla. Tavallinen tietokoneen kiintolevy toimii käyttöjärjestelmässä Block storagen kanssa identtisesti tiedostoja tallentaessa ja tiedostot ovat pilkottu palasiksi kiintolevyllä. Block storage toteutetaan pilvialustoissa tallennusalueverkkojen (SAN) avulla, jotka ovat yleisimpiä varastointiverkkoarkkitehtuureja. SAN yhdistää useita tallennuslaitteita yhdeksi verkoksi, johon tallennustilaa käyttävät laitteet myös liittyvät.

SAN mahdollistaa verkon luomisen muutamalla eri protokollalla, mutta kaikki prokollat käyttävät SCSI-protokollaa. SCSI on vanha, vuonna 1986 luotu protokolla, joka määrittää tiedonsiirtoväylän sekä mitenkä siihen kytketyt laitteet kommunikoivat keskenään. Useat tietokonearkkitehtuurit sisältävät erillisen väylän I/O operaatioille, joiden avulla CPU juttelee I/O laitteiden kanssa. Tämä väylä toteuttaa SCSI-protokollan. SCSI-väylällä voidaan yhdistää useita tallennuslaitteita ja palvelimia, mutta sen skaalautuvuus on erittäin heikko. Luotuun verkkoon voidaan lisätä vain rajallinen määrä laitteita, sekä kaapelin pituudella on maksimi pituus. Versiosta riippuen verkkoon voidaan lisätä 8-16 laitetta sekä laitteet voivat olla toisistaan vain noin 20 metrin päässä toisistaan. Koska SCSI-väylät eivät ole järkeviä tallennusalueverkkojen luomista varten ne on korvattu järkevimmillä väylillä.

Valokuituverkot käyttävät valokuituja tiedonsiirtoon ja Fibre Channel Protocol (FCP) on verkon protokolla. Valokuituverkot ottavat vastaan SCSI-komentoja ja muuttavat ne FCP-protokollan komennoiksi valokuiduissa tapahtuvan siirron ajaksi ja muuttavat ne takasin SCSI-komennoiksi kun siirto on tehty. Tätä varten tarvitsee asentaa oma laiteajurinsa käyttöjärjestelmään. Tallennustilaa käyttävät laitteet ja tallennuslaitteet kytketään samaan valokuituverkkoon, jolloin tiedostonsiirrot eivät rasita normaalia intenetyhteyttä. Tämä prokolla on eniten käytetty ja nopein, koska valokuituverkko on erittäin nopeaa tiedostojen siirtoon. Haittapuolena on verkon erittäin korkea hinta ja verkon ylläpitämiseen tarvittava erikoisosaaminen. Toisista yleisin protokolla on iSCSI, joka käyttää internetyhteyttä valokuituverkon sijasta, ja on huomattavasti halvempi kuin valokuituverkot, mutta hitaampi. iSCSI siirtää SCSI-komennnot TCP-protokollan avulla verkossa. iSCSI on myös tehottomampi tapa siirtää dataa, koska TCP-pyynnöissä on paljon enemmän turhaa dataa mukana suhteuttuna SCSI-komentoihin verrattuna FCP-protokollaan. Muita TCP:tä käyttäviä SAN-protokollia ovat: iFCP ja FICP, mutta ne ovat huomattavasti vähemmin käytetty mitä iSCSI.

Pilvipalveluiden tarjoamat Block storaget ovat lähes poikkeuksessa toteutettu iSCSI protokollan avulla ja ne voi sitoa virtuaalikoneisiin tavallisena kansioina. Block storageihin tallennettu data ei ole sidottu virtuaalikoneen elinkaareen ja data säilyy vaikka virtuaalikoneen poistaisi mihin Block storage oli sidottu.

### File storage

File storage tallentaa datan hierarkisessa rakenteessa käyttäen tiedostoja ja kansioita. File storage mahdollistaa levyn käytön usealle yhtäaikaiselle käyttäjälle ja sen toteutus perustuu samoihin teknologioihin kuin tavallisen verkkolevyn. Yleisin teknologia on NAS, joka on esitetty kuvassa 7. NAS käyttää sisäisesti samoja teknologioita kuin SAN, mutta se lisää oman protokollakerroksen niiden päälle hallitsemaan tiedostoja. NAS-palvelin avaa IP-osoitteen, mihin asiakkaan voivat ottaa yhteyden LAN-verkosta. Tällöin tiedostot liikkuvat samassa verkossa, missä muukin internetliikenne tapahtuu ja tiedostojen siirto voi rasittaa muuta internetliikennettä. NAS edellyttää asiakkailta erillistä asikasohjelmaa, joka on käyttöjärjestelmästä riippuvainen. Yleisin asikasjärjestelmän protokolla on NFS.



Kuva 10 NAS-arkkitehtuuri

## Pilvialustat

Tässä kappaleessa käsitellään projektissä läytettävät julkisen pilven palvelut.

### Microsoft Azure

Microsoft Azure on Microsoftin ylläpitämä pilvipalvelu, joka julkistettiin Microsoftin Professional Developers Conferencessa (PDC) lokakuussa 2008, ja se julkaistiin virallisesti helmikuussa 2010 nimellä Windows Azure vaihtoehdoksi silloin jo tunnetuille Amazon ja Google pilvialustoille. Azuren ensimmäinen versio tarjosi hyvin rajoitetun joukon palveluita, pelkästään pilvipalvelun ASP.NET-verkkosovellusten kehittämiseen ja käyttämiseen, Azure Blob -tietovaraston, Azure SQL -pilvitietokannan ja Azure Service Busin. Vuonna 2014 sen nimi muutettiin nykyiseksi Microsoft Azureksi. Alustan taustatarina alkoi vuonna 2005, kun Microsoft osti Groove Networksin ja Microsoftin kehittäjät alkoivat kehittämään pilvikäyttöjärjestelmää. Azurea koodattiin aluski koodinimella Red Dog ja se julkistettiin vuonna 2008. Se oli tuolloin Windows NT:n laajennus, joka oli suunniteltu toimimaan pilvessä. Se tarjoaa ohjelmistoa palveluna (SaaS), alustan palveluna (PaaS) ja infrastruktuurin palveluna (IaaS) ja tukee monia erilaisia ​​ohjelmointikieliä, työkaluja ja kehyksiä, mukaan lukien sekä Microsoft-kohtaiset että kolmannen osapuolen ohjelmistot ja järjestelmät.[ https://apix-drive.com/en/blog/reviews/microsoft-azure-review]

Azure, kuten muut pilvialustat perustuu virtualisointiteknologiaan. Tietojenkäsittelyssä virtualisointi tarkoittaa virtuaalisen eikä todellisen version luomista jostakin tietokoneen osasta, mukaan lukien virtuaaliset tietokonelaitteistot, tallennuslaitteet ja tietokoneverkkoresurssit.. Suurin osa tietokonelaitteistoista voidaan voidaan virtualisoida ohjelmistossa. Tietokonelaitteisto on yksinkertaisesti joukko ohjeita, jotka on pysyvästi tai puolipysyvästi koodattu piihin. virtualisointikerroksia käytetään yhdistämään ohjelmistoohjeet laitteiston ohjeisiin. virtualisointikerrokset mahdollistavat virtualisoidun laitteiston suorittamisen ohjelmistossa, kuten itse laitteisto. [https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/get-started/what-is-azure]

Pohjimmiltaan pilvialustat ovat joukko fyysisiä palvelimia yhdessä tai useammassa datakeskuksessa. Palvelinkeskukset toteuttavat virtualisoituja laitteistoja asiakkaille. Jokaisen palvelinkeskuksen sisällä on kokoelma palvelimia palvelintelineissä. Jokainen palvelinteline sisältää useita palvelinkortteja ja verkkokytkimen. Nämä tarjoavat verkkoyhteyden ja virranjakeluyksikön (PDU), joka tuottaa virtaa. Telineet on joskus ryhmitelty suuremmiksi yksiköiksi, jotka tunnetaan klustereina. Palvelimen telineet tai klusterit suorittaa virtualisoituja laitteisto-ilmentymiä käyttäjille. Jotkut palvelimet käyttävät kuitenkin pilvihallintaohjelmistoa, joka tunnetaan nimellä kuituohjain. kuituohjain on hajautettu sovellus, jolla on monia vastuita. Se allokoi palveluita, tarkkailee palvelimen ja siinä käynnissä olevien palveluiden kuntoa ja korjaa palvelimia, jos ne kaatuvat. Kukin kuituohjain on yhdistetty pilviorkesteriohjelmistoon jossa pyörii servereitä, Rajapintoja sekä sisäisiä tietokantoja joita Azure vaatii toimiakseen. pilviorkesteriohjelmistossa pyörii muunmuissa palveluita jotka vastaavat käyttäjien tekemiin pyyntöihin hallintakäyttöliittymän kautta. Käyttäjän pyynnöt allokoivat Azure-resursseja ja -palveluita. Ensin käyttöliittymä varmistaa, onko käyttäjällä lupa varata pyydetyt resurssit. Jos näin on, käyttöliittymä tarkistaa tietokannan löytääkseen riittävän kapasiteetin omaavan palvelintelineen, joka kehottaa kuituohjainta varaamaan resurssin [https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/get-started/what-is-azure]

## Tietoturva

Tässä kappaleessa käsitellään tietoturvaan liittyviä protokollia

### RSA

RSA (Rivest–Shamir–Adleman) on julkisen avaimen salausjärjestelmä, joka on yksi vanhimmista yleisesti käytetyistä salausjärjestelmistä. RSA salaa tai allekirjoittaa tiedot kahdella eri avaimella. Yksi avaimista on julkinen avain, joka on kaikkien käytettävissä ja toinen on salainen yksityinen avain. Julkisella avaimella salattujen tietojen salaus voidaan purkaa vain yksityisellä avaimella. Allekirjoittamisessa yksityistä avainta käytetään allekirjoituksen luomiseen ja kuka tahansa voi julkisella avaimella todentamaan että allekirjoitus on tehty yksityisellä avaimella. Kahden avaimen käytön vuoksi julkisen avaimen salaus tunnetaan myös epäsymmetrisenä salauksena, koska salausta ei pureta samalla avaimella millä se luotiin. Symmetrisessä salauksessa samaa avainta käytetään salauksen luomiseen ja purkamiseen.

Nimi RSA tulee Ron Rivestin, Adi Shamirin ja Leonard Adlemanin sukunimistä, jotka kuvailivat algoritmin julkisesti vuonna 1977. RSA:n turvallisuus perustuu käytännön vaikeuteen jakaa kahden suuren alkuluvun tulo tekijöihin. RSA-salauksen rikkominen tunnetaan RSA-ongelmana. Ongelmaa ei ole onnistuttu ratkaisemaan jos käytetään riittävän suurta avainta. RSA on suhteellisen hidas algoritmi. Tästä johtuen sitä ei yleensä käytetä suoraan tietojen salaamiseen, vaan sillä salataan jaettuja avaimia symmetristä salausta varten.

RSA:n perusperiaate on, että on helppoa löytää kolme erittäin suurta positiivista kokonaislukua , ja , siten että kaikille kokonaisluvuille pätee että jakojäännös jaettuna on . Matemaattisesti ilmaistuna kaavalla.

Kun annetaan vain ja , on erittäin vaikea löytää . Kaavassa Kokonaisluvut ja muodostavat julkisen avaimen, on yksityinen avain ja on viesti. Koska ja d voidaan vaihtaa kaavassa päikseen, mahdollistaa se viestien allekirjoittamisen ja allekirjoituksen tarkistamisen samaa algoritmia käyttäen. julkinen avain muodostetaan kahden satunnaisen suuren alkuluvun tulosta ja sen pituus on julkisen avaimen pituus. on positiivinen satunnainen kokonaisluku, jonka arvon valintaväli määritetään Carmichaelin funktion avulla. Yksityinen avain saadaan myös laskettua Carmichael funktion ja julkisen avaimen avulla. Viesti saadaan salattua laskemalla seuraava salakirjoitus käyttäen julkista avainta ja kaavalla

### TLS

TLS on Julkisen avaimen salausmenetelmä, joka on luotu vanhentuneen SSL-protokollan päivitetyksi versioksi. SSL:n ensimmäistä versiota ei koskaan julkaistu tietoturvaongelmien takia, mutta ensimmäinen julkaistu versio SSL-protokollasta oli SSL 2.0, joka julkaistiin vuonna 1995. SSL 2 löytyi heti tietoturvaongelmia ja ne ongelmat korjattiin sueraavana vuonna julkaistussa SSL 3.0 versiossa. TLS 1.0 julkaistiin vuonna 1999 ja siinä ei ole merkittäviä eroja SSL 3.0 versioon. Vuonna 2006 julkaistu TLS 1.1 lisäsi tietoturvaa. TLS 1.2 julkaistiin vuonna 2008 ja nykyinen uusin versio 1.3 julkaistiin vuonna 2018. Uddet versiot ovat parantaneet tietoturvaa poistamalla protokollasta murtuneita salausalgoritmeja ja pakottaneet käyttämään salausalgoritmeissa vahvempia avaimia.

TLS käyttää X.509 -sertifikaattistandardin varmenteita viestintäkumppanin todentamiseen. Kun asikas ottaa yhteyden alvelimeen, palvelin lähettää varmenteen asiakkaalle. Varmenteessa olevat tärkeät tiedot ovat verkkotunnuksen nimi jolle varmenne on myönnetty, Varmenneviranomainen joka on myöntänyt varmenteen ja sen luoma allekirjoitus, voimassaoloaika sekä palvelimen oma julkinen avain. Varmenneviranomainen (CA) on toimija, joka tallentaa, allekirjoittaa ja myöntää digitaalisia varmenteita. Se on kolmannen osapuolen palvelu, johon molemmat osapuolet luottavat. Asiakkaalla on lista Varmenneviranomaisista joihin se luottaa ja niiden julkisista avaimista. Luotetun julkisen avaimen avulla se pystyy varmentamaan palvelimen varmenteen varmenneviranomaisen allekirjoituksen joka on tehty varmenneviranomaisen yksityisellä avaimella. Varmenneviranomaisen tehtävä on varmistaa ennen palvelimen varmenteen allekirjoittamista että palvelin omistaa kysyisen verkkotunnuksen, jolle se on pyytämässä allekirjoitusta. Verkkotunnuksen varmentaminen onnistuu esimerkiksi DNS-tietueiden avulla. Varmenneviranomaisen tekemän allekirjoituksen avulla asiakas pystyy luottamaan palvelimen identiteettiin. Varmenne voidaan myös allekirjoittaa omalla yksityisellä avaimella, jolloin puhutaan itseallekirjoitetuista varmenteista. Silloin asiakas ei voi varmentaa palvelinta varmenneviranomaisten avulla vann luottamus pitää todentaa toisella tapaa. <https://aws.amazon.com/what-is/ssl-certificate/> https://www.ssl.com/article/ssl-tls-self-signed-certificates/

TLS-kättely on prosessi, joka käynnistää TLS:ää käyttävän viestintäistunnon. TLS-kättelyn aikana kaksi kommunikoivaa osapuolta vaihtavat viestejä vahvistaakseen toisensa, sopiakseen käytettävät salausalgoritmit ja sopiakseen istuntoavaimista. TLS versioissa ennen 1.3 versiota ensimmäisenä asiakas lähettää viestin palvelimelle, jossa se ilmoittaa tukemansa TLS versiot ja salaukset. Lisäksi asiakas lähettää satunnais generoidun numeron. Palvelin vastaa tähän viestiin korkeimman TLS version mitä molemmat osapuolet tukevat sekä molempien tukeman salauksen. Palvelin lähettää oman satunnais generoidun numeron sekä SSL-sertifikaatin asiakkaalle, jonka avulla asiakas varmistaa sertifikaatin aitouden. Kun aitous on varmistettu, asiakkaalla on useampi mahdollinen tapa jakaa symmetriseen salaukseen käytettävä avain palvelimen kanssa. Yksinkertaisin on RSA- kättely, jossa asiakas lähettää palvelimelle avaimen salaamalla sen palvelimen julkisella avaimella. Näin palvelin pelkästään pystyy purkamaan salauksen sen omalla yksityisellä avaimellaan. Monimutkaisempi tapa avaimen vaihtoon on käyttää Diffie–Hellman-avaintenvaihtoprotokollaa (DH). DH-protokollassa avainta ei lähetetä salattuna toiselle osapuolelle, vaan molemmat osapuolet laskevat salaisuuden julkisia arvoja käyttäen. DH- protokollassa valitaan alkuluku ja sen generaattori . Molemmat luvut jaetaan julkisesti osapuolten välillä. Molemmat osapuolet valitsevat itselleen salaisen luvun ja laskevat siitä julkisen luvun . Osapuolet jakavat julkiset luvut toisilleen ja laskevat siitä yhteisen salaisuuden kaavlla , jossa on toisen osapuolen julkaisema julkinen luku ja on oma salainen luku. Salaisuuden jakamiseen voidaan myös käyttää DH-protokollasta johdettuja monimutkaisempia algoritmeja, joissa käytetään elliptisiä käyriä ja niiden krptografisia ominaisuuksia. Kun salaisuus on jaettu, asiakas lähettää salatun viestin palvelimelle ja palvelin vastaa siihen salatun viestin. Tämän jälkeen kättelyosuus on valmis ja yhteys jatkuu symmetrisellä salauksella salattuna, jossa salaukseen käytetään palvelimen ja asiakkaan lähettämiä satunnaisia lukuja ja yhteistä salaisuutta.

TLS 1.3 yksinkertaistaa kättelyosuutta merkittäväst sekä parantaa tetoturvaa poistamalla epäturvallisia algoritmeja protokollasta. Yhteisen salaisuuden jakamiseen 1.3 versio edellyttää eteenpäin turvallisuutta. Tämä tarkoittaa jos palvelimen yksityinen salaisuus paljastuu tulevaisuudessa, niin sillä ei pysty purkamaan sillä tehtyjä vanhoja salauksia. Tämän takia edellisessä kappaleessa esitetty RSA- kättely ei ole mahdollista ja DH-protokollassa edellytetään jokaiselle yhteydelle käytettäväksi aina uusia julksia parametrejä. Asiakas lähettää ensimmäisessä viestissä satunnaisgeneroidun numeron, sen tukemat symmetrisen saluksen algoritmit sekä valitsee menetelmän yhteisen salaisuuden jakamiseen ja laskee siihen tarvittavat julkiset luvut. Asiakas ei voi tietää tässä vaiheessa että palvelin tukee valittua menetelmää yhteisen salaisuuden jakamiseen, mutta vaihtoehtoja on 1.3 versiossa vähemmän joten palvelin tukee asiakkaan valitsema mentelmä erittäin todennäisesti. Palvelin vastaa tähän viestiin valitulla symmetrisellä salauksella, omalla satunnaisluvulla sekä valitulla menetelmällä salaisuuden vaihtoon johon se on laskenut omat julkiset arvonsa. Palvelin pystyy tässä vaiheessa laskemaan yhteisen salaisuuden ja on valmis symmetriseen salaukseen. Palvelimen viesti toimii samalla kuittauksena aloittaa salattu yhteys. Asiakas laskee jaetun salaisuuden palvelin vastauksen parusteella ja on valmis symmetriseen salaukseen. Asiakas varmistaa palvelimen varmenteen aittouden ja lähettää kuittausviestin palvelimelle. Kättely on suoritettu ja yhteys jatketaan symmetrisellä salauksella. TLS 1.3 tukee myös TLS-kättelyn entistä nopeampaa versiota, joka ei vaadi ollenkaan edestakaisia viestejä. Jos asiakas ja palvelin ovat muodostaneet yhteyden toisiinsa aiemmin, ne voivat luoda toisen jaetun salaisuuden ensimmäisessä yhteydessä yhteyden jatkamista varten. Palvelin lähettää asiakkaalle myös istuntolipun yhteyden aikana. Seuraavan yhteyden alkaessa asiakas voi käyttää istuntolippua ja yhteyden jatkamiseen luotua salaisuutta ja TLS-salaus pystytään luomaan ilman kättelyvaihetta.

<https://www.comparitech.com/blog/information-security/diffie-hellman-key-exchange/> <https://www.upguard.com/blog/diffie-hellman>

https://www.thesslstore.com/blog/tls-1-3-handshake-tls-1-2/

Symmetriseen salaukseen käyetetään joko lohko- tai virtasalausalgoritmia. Lohkosalausalgoritmi ottaa vakiokokoisen lohkon kokoisen määrän dataa ja salaa sen salatekstiksi. Virtasalauksessa jokainen salattavan tekstin bitti salataan yksi kerrallaan salausnumerovirran vastaavalla bitillä, jotta saadaan salatekstivirran bitti. Lohkosalusta voidaan käyttää virtasalauksen tapaan jos käytetään alustusvektoreita jotka salataan lohkosalauksella ja salattu alustusvektori yhdistetään salattavaan tekstiin.TLS:sän symmetrisessä salauksessa käytetään lohkosalausta virtausalgoritmin tapaan. Salauksen lisäksi algoritmin pitää pystyä todentamaan salatun datan aitous, eli että kukaan ei ole muokannut salattua dataa siirron aikana. Tälläisiä salausjärjestelmiä, jotka pystyvät salaaman datan ja todentamaan salatun data aitouden kutsutaan todennetuksi salaukseksi (AE). Yleisin AE- algoritmi on GCM (Galois/Counter Mode), jota TLS- protokolla käyttää. TLS 1.3 mahdollistaa myös toisen EA- algoritmin käyttämisen CCM (Counter with CBC-MAC), jota aikaisemmat versiot eivät tue.

TLS 1.2 ja aikaisemmat versiot käyttävät useita eri lohko- ja virtasalausalgoritmeja, mutta 1.3 versio käyttää vain yhtä lohkosalausta (AES) ja yhtä virtasalausta (ChaCha20). AES, joka tunnetaan myös alkuperäisellä nimellä Rijndael on Yhdysvaltain kansallisen standardisointi- ja teknologiainstituutin (NIST) vuonna 2008 vahvistama spesifikaatio lohkosalaukseen. AES on nykyään yksi käytetyimmistä lohkosalausalgoritmeista ja siitä on olemassa variaatiot 128, 192  ja 256 bittisille avaimille. Kuvassa 9 on esitetty lohkokaavio AES-GCM todennetun salauksen järjestelmästä, jota TLS 1.3 ja TLS 1.2 käyttää.

A diagram of a computer program

Description automatically generated

Kuva 11 AES-GCM

Lohkokaavion yläosassa lasketaan salattu data AES- algoritmilla ja alaosassa lasketaan varmenne salatusta datasta, jonka avulla pytsytään todentamaan että salattua dataa ei ole muokattu siirron aikana. Salauksessa käytetään alustusvektoria (IV) ja siihen lisätään lohkonumero ja tämä summa salataan AES:in avulla. Varmenne lasketaan Galois'n kuntien avulla ja siihen saadaan luotua avain AES-salaukseen käytetyn avaimen avulla. Varmenteen laskemiseen käytetään salattua dataa, alustusvektoria sekä mahdollista ylimääräistä dataa. AES-GCM mahdollistaa myös ylimääräisen salaamattoman datan lähettämisen salatun datan kanssa ja pystyy myös varmentamaan tämän aitouden ottamalla sen mukaan vermenteen laskemiseen. Tämä on merkattu kaavioon AAD. <https://datatracker.ietf.org/doc/html/rfc8446> <https://csrc.nist.rip/groups/ST/toolkit/BCM/documents/proposedmodes/gcm/gcm-spec.pdf> <https://infocenter.nordicsemi.com/index.jsp?topic=%2Fcom.nordic.infocenter.nrf52832.ps.v1.1%2Fccm.html>

Kuva 10 esittää F5 Labsin tekemän The 2021 TLS Telemetry Report- tutkimuksen tuloksia TLS:n käytetyimmistä salausalgoritmeista. Mukana tutkimuksessa on miljoona käytetyintä nettisivustoa. Salausalgoritmit ovat iloitettu kuvassa muodossa yhteisen salaisuuden jakaminen-symmetrinensalaus-MAC. Tutkimuksen mukaan neljä yleisintä salausta käyttävät kaikki symmetriseen salaukseen AES-GCM- algoritmia, joka esiteltiin edellisessä kappaleessa. Kun lasketaan neljän suosituimman algoritmin käyttöprosentit yhteen saadaan tulokseksi että AES-GCM suojaa 93.2 % kaikista TLS- yhteyksistä. Lisäksi kuvasta huomataan, että yhteisen salaisuuden jakamiseen käytetään DH-protokollan monimutkaisempaa elliptisiin käyriin perustuvaa versiota, joka on merkattu kuvaan lyhenteellä ECDHE. https://www.f5.com/labs/articles/threat-intelligence/the-2021-tls-telemetry-report

A screenshot of a computer

Description automatically generated

Kuva 12 Suosituimmat TLS- salausalgoritmit vuonna 2021

# Projektin esittely

Tässä kappaleessa esitellään anonyymisti projekti, johon diplomityön käytännön osuus perustuu. Projekti on anomisoitu asiakkaan toiveen mukaan, joten sitä ei tulla mainitsemaan nimellä. Projektin päämääränä on toteuttaa pilvipalvelu, joka mahdollistaa suurien tiedostojen lataamisen ja niiden automaattisen käsittelemisen ja tulosten siirtämisen eri tietovarastoihin. Projektiin valittiin Oracle Cloud -pilvipalvelu asiakkaan toiveesta. projekti on toteutettu mikropalvelu -arkkitehtuurina käyttäen Docker kontteja, joita ajetaan tuotannossa Oracle Cloudin ylläpitämässä Kubernetes -palvelussa .

Tiedostojen lataamiseen valittiin ratkaisuksi selain pohjainen ratkaisu, mutta myös muita vaihtoehtoja tutkittiin. Ohjelman pitää pystyä lataamaan tiedostoja sekä kansioita ja sen pitää pystyä validoimaan, että tiedostot eivät ole korruptoituneet siirron aikana. Latauksen suuruus on luokkaa 1 – 10 Gb, ja lataus pitää pystyä tekemään normaalitehoisella loppukäyttäjän tietokoneella.

Tiedostot ladataan jaettavalle levylle, joka voidaan jakaa useampien konttien kesken. Latauksen jälkeen toinen kontti käsittelee ladattavat tiedostot ja siirtää tulokset eri tietovarastoihin. Käytettäviä tietovastoja on SQL -tietokannat, jaettavat verkkolevyt sekä Object Storage.

Valitut teknologiat valittiin arvioimalla niiden suorituskykyä, hintaa sekä tietoturvallisuutta. Tässä diplomityössä käytännönosuutena on toteutettu pieniä testipenkkejä, jotka mittaavat vaadittuja omanaisuuksia, sekä on vertailtu eri pilvialustojen hintoja.

Seuraavassa luvussa 4 Tulokset käytetään seuraavia ominaisuuksia vertaillessa lataus- ja talletustapoja. Tietoturva, nopeus, käytetty verkkoyhteys latauksen yhteydessä sekä kustannukset.

Tietoturva käsitellään kootusti yhtenä kappaleena jossa vertaillaan käytettyjen protokollien tietoturvallisuutta ja mahdollisia haavoittuvuuksia käytetyissä teknologioissa.

# TULOKSET

Työssä esitellään useampi pienempi testausprojekti, jotka voidaan jakaa karkeasti kahteen osioon: tiedostojen siirtoon sekä tiedostojen talletukseen. Kokonaisjärjestelmän arkkitehtuuri on esitetty kuvassa 8, joka kattaa sekä tiedostojen siirron että talletuksen. Järjestelmän luomiseen on käytetty Azure-pilvipalvelua ja testejä ajetaan Azuren ulkopuolella paikallisella tietokoneella. Azuressa on käytössä virtuaalikone, jossa pyörii .NET:n avulla toteuttu palvelin, Nginx palvelin, joka pyörittää käyttöliittymää, jota käytetään HTTP-latausten testaamiseen. Käyttöliittymä on toteutettu Angular ohjelmointikehyksen avulla ja se on toteutettu Single Page Application (SPA) arkkitehtuurilla. Nginx-palvelinta käytetään SPA-käyttöliittymän jakamiseen internettiin. Virtuaalikoneella on myös asennettu vsftp-palvelin, joka on FTP palvelin, joka mahdollistaa myös FTPS-yhteyden. OpenSSH-palvelinta käytetään SFTP-yhteyden luomiseen. Virtuaalikoneeseen on liitetty Bock ja File storaget ja ne on sidottu osaksi virtuaalikoneen tiedostojärjestelmää ja niihin pääse käsiksi samaan tapaan kuin virtuaalikoneen omaan kovalevyyn. Blob storageen ja Dynamo-tietokantaan pääsee käsiksi julkisen HTTP rajapinnan kautta. Dynamo-tietokanta on Azuren toteuttama NoSQL-tietokanta, jossa data on tallennettuna json-dokumentteina. Virtuaalikoneeen kanssa samassa virtuaalisessa sisäverkossa on myös Postgres SQL-palvelin. Virtaalikone juttelee SQL palvelimen kanssa C# ohjelmoiintikielellä toteutetun asiakasohjelman avulla.

Diagram

Description automatically generated

Kuva 13 Testiohjelman kokonaisarkkitehtuuri

## Tiedostojen siirto

Tiedostojen siirtoon liittyvät projektit ovat jaoteltu useampaan pienempään projektiin eri tiedostonsiirtotapojen perusteella. FTP-, FTPS- ja SFTP-siirrot muodostavat yhden projektin ja selaimella toimivat protokollat muodostavat oman projektin.

FTP, FTPS ja SFTP siirrot ovat toteutettu kahdella eri ohjelmointikielen kirjastolla. Käytetyt ohjelmointikielet ovat Python sekä C#. Python-ohjelmointikielelllä on FTP ja FTPS kirjasto on Pythonin sisäinen kirjasto siihen ei tarvita kolmannen osapuolen kirjastoja. SFTP-yhteys on luotu Paramiko-kirjastolla, joka on kolmannen osapuolen ilmaisen lähdekoodin kirjasto. C#-koodissa on läytetty kolmannen osapuolen Nuget-paketteja yhteyksien luomiseen. FluentFTP-kirjastoa on käytetty FTP- ja FTPS- yhteyksien luomiseen ja SSH.NET-kirjastoa SFTP-yhteyteen. Projektissa on ladattu erikokoisia tiedostoja käyttäen kumpaakin ohjelmointikieltä. Tiedostojen koot ovat olleet kilotavun ja 10 megatavun välissä. Saman kokoinen tiedosto on ladattu viiteen kertaan ja jokaisen latauksen latausaika on kirjattu muistiin. Näistä latausajoista on laskettu keskiarvo kullekin tiedostokoolle ja keskiarvoja on käytetty tulosten käsittelemiseen

Selaimella toimivia protokollia eli HTTP- ja WebSocket-prokollia on testattu selainpohjaisen käyttöliittymän avulla. Kuvassa 9 on näyttökuva käyttöliittymän HTTP-latauksesta

Graphical user interface, text, application

Description automatically generated

Kuva 14 HTTP-latausnäkymä

Käyttöliittymällä pystyy valitsemaan eri asetuksia käytettäviin protokolliin. HTTP-latauksessa tiedostoa ei lueta kokonaan selaimeen muistiin, vaan se luetaan palasina. Jos tiedostoa ei palotelttaisi, koko tiedosto pitäisi lukea selaimen käyttämään RAM-muistiin ja se rajoittaisi merkittävästi lähetettävän tiedoston maksimikokoa. Tiedoston paloittelemisen avulla lähetettävällä tiedostolla ei ole maksimikokoa rajoittavaa tekijää. Käyttöliittymässä on valittavana 5kB-, 50kB- ja 500kB-kokoiset palaset. Kukin palanen läheteään omana POST pyyntönä palvelimellle ja viimeisen palasen jälkeen palvelin kokoaa palaset yhteen ja muodostaa kokonaisen tiedoston palvelimella.Toinen valittava ominaisuus on lähetettävän palasen koodaus. Kaikissa koodauksissa tiedoston palanen luetaan selaimen File API:n avulla ArrayBuffer-tyyppisenä selaimen muistiin. Arraybuffer on datan esitys kahdeksan bittisenä numeroina taulukkotietorakenteessa. Data voidaan siirtää käyttäen eri koodauksia siirrossa. Base64-koodauksessa binäärinen data muunnetaan merkkijonoksi. Base64 koodisto käyttää 64-merkkistä aakkostoa johon binääridata muunnetaan. Number array-valinta lähettää binäärisen datan merkkijonona ilman koodausta. Tämä vaihtoehto kasvattaa huomattavasti siirrettävän tiedoston kokoa, koska jokaisen tavun jokainen numero esitetään UTF-8 merkkinä siirron aikana. Base64 ja Number array siirtävät datan json-formaattina, jonka merkistö on UTF-8. Form-data formaatti siirtää datan binäärimuodossa ja dataa ei tarvitse lähettää merkkijonona.

WebSocket-protokollalla on mahdollista valita siirrettävän datan formaatti. Valittavana on merkkijono sekä binääri. Merkkijono käyttää UTF-merkistöä siirrettävän datan koodauksena ja binääri käyttää binääriformaattia tiedonsiirtoon.

Käyttöliittymää ohjataan testeissä Selenium-kirjastolla, jolla pystyy automatisoimaan selaimen käyttöä. Seleniumin avulla selainta ohjataan toistamaan aina samat valinnat. Kun tiedosto on ladattu onnistuneesti selaimeen tulee latausaika näkyviin, jonka Selenium poimii talteen.

## Tiedostojen talletus

Tiedostojen talletukseen on toteutettu .NET-palvelin joka suorittaa talletusoperaatiot sekä Python ohjelma joka tekee pyynnön .NET-palvelimelle ja muodostaa tuloksista csv-tiedoston sekä kuvaajan.

# Projektien tulokset

Projektien tulokset ovat jaettu suorituskyvyn sekä kustannusten mukaan. Kustannusvertailussa otetaan kantaa talletusratkaisujen kustannuksiin Azure-pilvipalvelussa ja vertaillaan eri talletusratkaisujen kustannuksia. Tehokkuusvertailussa vertaillaan talletus- sekä lataustapoja ajallisessa näkökulmassa sekä lataustapojen käyttämää verkon määrää.

## Kustannusvertailu

Kustannusvertailu on jaettu kahteen osaan: Pilvialustojen talletusratkaisuihin ja tiedostojen siirtoon. Tiedostojen siirron kustannuksia mitataan epäsuorasti kuinka paljon siirrettävän tiedosotn koko kasvaa siirron aikana eri protokollilla. Siirrettävän datan määrä voi vaikuttaa internetyhteyden hintaan sekä pilvipalvelut laskuttavat virtuaalikoneille tulevasta ja lähtevästä datasta.

Tutkimus tehtiin käyttämällä WireShark-ohjelmistoa, joka avoimen lähdekoodin pakettianalysaattori. Wireshark on monialustainen, ja se käyttää pcap:ia pakettien sieppaamiseen.

### Pilvialustojen talletusratkaisujen kustannukset

Tässä osiossa vertailtaan Azuren tarjoamien talletusratkaisujen hintoja.

Azure Cosmos on NOSQL tietokanta sekä vektoritietokanta, joka tallentaa dokumentteja JSON – muodossa. Dokumentit indeksoidaan ja niitä on mahdollista hakea ja muokata kyselyillä. Azure Cosmos laskuttaa kolmesta eri käyttötyypistä: laskenta, tallennustila ja kaistanleveys. Azure tarjoaa ilmaiset 25 GB talletustilaa jokaiselle azuren tilille. Laskennan laskutus tapahtuu pyyntöyksikköjen avulla (RU). RU on tietokannan operaatioihin vaadittavan laskennan, muistin ja IO:n abstrakti mitta. eri tietokantaoperaatiot kuluttavat pyyntöyksikköjä tietyn välin sisällä. Tietyn operaation vaadittavaan pyyntöyksikköjen määrään vaikuttaa useampi tekijä esimerkiksi kirjoitettavan/luettavan asian koko sekä indeksien käyttäminen kasvattaa vaadittavien Pyyntöyksikköjen määrää. Vaadittavien pyyntöyksikköjen määrä on deterministinen, eli sama operaatio, samalla tietokannalla vaatii aina saman verran Pyyntöyksikköjä. Azure Cosmos DB laskuttaa pyyntöyksiköillä (RU) sekunnissa mitattuna (RU/s) tai pelkästään käytetyistä pyyntöyksiköistä. Laskutus riippuu tietokannan valitusta skaalautuvuusmallista. palveliton malli laskuttaa pelkästään käytetyistä pyyntöyksiköistä 0.235 euroa miljoonasta pyyntöyksiköstä. Kiinteäsuorituskyky-mallissa tietokantaan kiinnitetään kiinteä RU/s arvo, ja siitä veloitetaan 0.0075 euroa tunnilta sataa RU/s yksikköä kohden. Automaattisessa skaalaus- mallissa tietokannan RU/s arvo skaalautuu automaattisesti kuorman mukana 10% -100% välillä annetusta maksimi RU/s arvosta. Tunnin maksimi RU/s arvosta veloitetaan 0.008 euroa tunnilta sataa RU/s yksikköä kohden. Talletustila maksaa 0.235 euroa kuukaudessa giga tavulta. Kaistanleveydestä joutuu maksamaan pelkästään tietokannasta ulos siirtyvästä datasta. Ensimmäiset 5 GB on ilmaista, jonka jälkeen hinta on 0.05 euroa giga tavulta. https://azure.microsoft.com/en-us/pricing/details/cosmos-db/autoscale-provisioned/

<https://learn.microsoft.com/en-us/azure/cosmos-db/request-units>

Azure tarjoaa täysin ylläpidetyn SQL tietokannan ja varmistaa sen saatavuuden ja suorituskyvyn. Tietokannoissa maksaa laskenta ja talletustila. laskentaan on tarjolla kaksi eri mallia: virtuaaliytimeen (vCore) ja Tietokantatapahtumayksikköön (DTU) perustuva ostomalli. Virtuaaliytimeen (vCore) perustuva malli tarjoaa valinnan varattujen tai palvelimettomien laskentamallien välillä. Valmistetun laskentamallin avulla voi valita tarkan määrän virtuaaliytimiä, jotka on aina varattu laskentaa varten. Palvelimettomassa laskentamallissa virtuaaliytimiä voidaan automaattisesti skaalata konfiguroitavalla laskenta-alueella. Palvelimeton laskentamalli sulkee automaattisesti tietokannan epäaktiivisina aikoina, jolloin tietokantaan ei ole yhtään istuntoa avattuna. Suljetusta ajasta laskennasta ei laskuteta olleenkaan, ja tietokanta käynnistyy udelleen kun uusi istunto luodaan. Virtuaaliytimiä on tarjolla eri arkkitehtuurilla ja niillä on hieman eri hinnat. Virtuaaliytimen määrä on valittavissa 2 ja 128 väliltä ja yhden virtuaaliytimen hinta pienenee hieman kun valitaan suurempi määrä ytimiä. Yhden ytimen hinta kuukaudessa on 100-120 euroa. Palvelimeton malli laskuttaa ytimestä sekuntihintaa ja varattu malli puolestaan minuuttihintaa. Tietokantatapahtumayksikkö (DTU) edustaa suorittimen, muistin, lukujen ja kirjoitusten abstraktia mittaa. DTU-pohjaisessa ostomallissa on mahdollista valita laskentatehon määrä joka ilmoitetaan käytettävissä olevilla tietokantatapahtumayksiköillä. Valittuun laskentatehoon on kiinnitetty kiinteä määrä mukana olevaa tallennustilaa, kiinteä säilytysaika varmuuskopioille ja kiinteä hinta. Lisämuistia on mahdollista ostaa hintaan kuuluvan muistin lisäksi. Laskentatehoa voidaan lisätä tietokannan luomisen jälkeen lisäämällä tietokantatapahtumayksikköjä ja se aiheuttaa vain lyhyen katkoksen tietokannan käyttämiseen. DTU-pohjaisen laskentatehon voi muuttaa virtuaaliytimen tehoon suhteella 100 DTU-yksikköä vastaa yhtä virtuaaliydintä.

Laskentateho on valittavissa 5-400 DTU-yksikön välistä ja käytettävissä oleva talletustila vaihtelee 2 GB - 4 TB välissä. Hinta taas vaihtelee 4.5 ja 14515 euron välillä. Virtuaaliytimeen perustuva malli laskuttaa talletustilasta 0.232 euroa kuukaudessa giga tavulta. Tietokantatapahtumayksikköön perustuvassa mallissa hintaan sisältyy tietty määrä levytilaa ja lisälevytila maksaa 0.16 euroa kuukaudessa giga tavulta https://learn.microsoft.com/en-us/azure/azure-sql/database/dtu-benchmark?view=azuresql

https://learn.microsoft.com/en-us/azure/azure-sql/database/resource-limits-dtu-single-databases?view=azuresql

https://azure.microsoft.com/en-us/pricing/details/azure-sql-database/single/

Azure Blob Storage on Microsoftin Object Storage ja se on suunniteltu rakenteettoman datan tallentamiseen. Azure Blob Storage tarjoaa HTTP-rajapinnan, jonka avulla tiedostoja pystytään muokkaamaan. Useille ohjelmistokielille on tehty valmiit asiakaskirjastot sekä CLI työkaluja. Blob Storage laskuttaa talletustilasta sekä operaatioista. Talletustilan ja operaatioiden hinta määräytyy valitun palvelumallin mukaan. valittavia malleja on viisi: korkealuokkainen, kuuma, lämmin, kylmä ja arkisto. Tallennuksen gigatavujen hinnat nousevat arkistosta kohti korkealuokkaista jokaisella askeleella. Puolestaan operaatiot halpenevat arkistosta kohti korkealuokkaista jokaisella askeleella. Korkealuokkaisella mallilla on pienin latenssi ja se kasvaa kohti arkistoa. korkealuokkaisenmallin talletus tila maksaa 0.13869 euroa giga tavulta ja on melkein kymmenen kertaa kalliimpi kuin seuraava kuumamalli. Hinnat halpenevat edelleen kohti arkistoa mentäessä ja arkisto maksaa 0.00092 euroa giga tavulta. korkealuokkainen malli tarjoa halvimmat luku- ja kirjoitusoperaatiot. korkealuokkaisen kirjoitusoperaatioiden hinta on 0.0211 € kymmentä tuhatta operaatiota kohden ja on noin kuusi kertaa halvempi kuin arkistomallin hinta. Lukuoperaatioiden hinta puolestaan muuttuu huomattavasti enemmän mallien välillä. korkealuokkaisen hinta on 0.0017 € kymmentä tuhatta operaatiota kohden. Kylmän mallin hinta on noin 100 kertainen korkealuokkaiseen nähden ja arkiston noin 4000 kertaa kalliimpi.

<https://azure.microsoft.com/en-us/products/storage/blobs/> <https://azure.microsoft.com/en-us/pricing/details/storage/blobs/>

Azure Files on Microsoftin File Storage palvelu. Palvelu mahdollistaa tiedostojen yhtäaikaisen käyttämisen eri käyttöjärjestelmien välillä. Valittavia palvelumalleja on neljä: korkealuokkaien, Tapahtuma optimoitu, kuuma, viileä. Korkealuokkaien tarjoaa pienimmän latenssin ja se kasvaa kohti kylmää mallia. Korkealuokkaisella talletustila maksaa eniten ja se halpenee mentäessä kohti kylmää. Korkealuokkaisen hinta on 0.1480 € varatulta giga tavulta ja kylmän mallin hinta on noin 10 kertaa halvempi. Luku- ja kirjoitusoperaatiot ovat korkealuokkaisessa ilmaisia, mutta muissa malleissa maksullisia. Tapahtuma optimoidussa on halvimmat hinnat ja hinnat kasvavat kun siirrytään kylmää mallia kohti. Kylmän mallin hinnat ovat keskimäärin noin kymmenen kertaa kalliimmat kuin Tapahtuma optimoidun. <https://azure.microsoft.com/en-us/pricing/details/storage/files/>

Azure Managed Disks on Microsoftin Block storage, jota voidaan käyttää virtuaalikoneiden kanssa. Levyvaihtoehtoja on neljä: Ultra Disk, korkealuokkaien SSD, SSD ja HDD. Ultra Diskin latenssi on kaikista pienin ja se on uuden generation SSD- levy. Korkealuokkaisen SSD:n ja SSD:n välillä on pieni suorituskyky- ja hintaero. Levyille on ilmoitettu kuinka monta kirjoitus/luku- operaatiota voidaan suorittaa (IOPS) sekunnissa sekä nopeus, jolla data voidaan siirtää ulos tai sisään, ilmoitettuna MB sekunnissa (MB/s). Azure tarjoaa mahdollisuuden parantaa levyn IOPS- ja MB/s-suorituskykyä hetkellisesti. Tätä kutsutaan purskeeksi ja maksimi yhtäkestoinen aika on sille 30 minuuttia. Purskeen käyttö kuluttaa krediittejä ja kun kaikki krediitit on käytetty purske loppuu. Kun levyä käytetään alle sen IOPS- ja MB/s-suorituskyvyn krediittejä kerääntyy uutta pursketta varten. Pienillä alle 128 GB levyillä korkealuokkisen SSD:n suorituskyky on huonompi kuin tavallisen SSD:n, mutta purske on paljon tehokkaampi. Mitä suurempi levyn koko sitä tehokkaampi korkealuokkainen SSD on verrattuna vastaavan kokoiseen tavalliseen SSD levyyn. Korkealuokkaisen SSD levyn hinta on noin 2 kertaa suurempi kuin vastaavan kokoisen SSD-levyn. Korkealuokkainen SSD levy laskuttaa vielä erikseen IO operaatioista 0.001850 € jokaista 10,000 operaatiota kohden tunnissa. IO-operaatioilla on maksimi hinta kuukaudessa ja se on noin levyn talletuskapasiteetin hinta. HDD-levyn hinta on noin puolet vastaavan kokoiseen SSD-levyyn nähden. HDD-levyn IOPS-suorituskyky on melkein sama kuin vastaavan kokoisen HDD-levyn, mutta siirrettävän datan määrä on huomattavasti pienempi. Ultra diskin laskutus toimii pikkaisen erillä lailla, siinä käyttäjä voi valita vapaasti halutun IOPS- ja MB/s-suorituskyvyn ja levyn talletuskoon. Jokaisesta valitusta komponentista laskutetaan erikseen ja levyn kustannukseksi tulee näiden summa. <https://azure.microsoft.com/en-us/pricing/details/managed-disks/> <https://learn.microsoft.com/en-us/azure/virtual-machines/disk-bursting>

Eri talletusratkaisujen hintoja vertaillaan kolmella eri talletuskoolla: 16 GB, 256 GB ja 1 TB. Talletusratkaisujen pelkästään talletustilasta laskuttaama hinta kuukaudessa on esitetty taulukossa 1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 16 GB | 256 GB | 1 TB |
| Managed Disks/ korkealuokkainen SSD | 2.22€ | 31.96€ | 113.61€ |
| Managed Disks/ SSD | 1.11€ | 17.76€ | 71.01€ |
| Managed Disks/ HDD | 0.70€ | 10.48€ | 37.87€ |
| Files/ korkealuokkainen | 2.37€ | 37.89€ | 148€ |
| Files/ tapahtuma optimoitu | 0.88€ | 14.2€ | 55.5€ |
| Files/ viileä | 0.22€ | 3.55€ | 13.9€ |
| Blob Storage/ korkealuokkainen | 2.22€ | 35.50€ | 138.69€ |
| Blob Storage/ kuuma | 0.27€ | 4.28€ | 16.7€ |
| Blob Storage/ arkisto | 0.01€ | 0.24€ | 0.92€ |
| Cosmos | 3.71€ | 59.40€ | 232€ |
| SQL | 3.71€ | 59.40€ | 232€ |

Taulukosta 1 näkee että tietokantojen (Cosmos ja SQL) hinnat ovat samat ja selvästi muita ratkaisuja kalliimmat. Kaikkien muiden talleyusratkaisuiden kalliimpien ratkaisuiden hinnat ovat keskenään lähellä toisiaan. Halvempien ratkaisuiden hinnat eroavat toisistaan enemmän. Kaikista halvin ratkaisu on Blob Storagen arkisto. Sen hinta verrattuna halvimpiin Managed Disksiin tai Files palveluihin on useita kymmeniä kertoja halvempi. Blob Storagen arkiston IO- operaatiot puolestaan ovat kymmeniä kertoja kalliimmat kuin minkään muun talletusratkaisun, eli se sopii parhaiten tiedostoille, joita ei tarvitse juuri olleenkaan muokata tai lukea.

Talletusratkaisuiden IO-kuormasta syntyviä kustannuksia, on paljon vaikeampi vertailla keskenään, koska talletusratkaisuilla on käytössä eri mittoja operaatioiden vaadittavan suorituskyvyn mittaamiseen. Taulukossa 2 on esitetty eri talletusratkaisuiden hinta miljoonalle lukuoperaatiolle ja miljoonalle kirjoitusoperaatiolle. Joillekkin talleyusratkaisulle on jouduttu arvioimaan hinta, koska talletusratkaisu ei suoraan ilmoita hintaa yksittäisille operaatiolle. Arvioidun hinnan edessä on käytetty tilde-merkkiä(~).

|  |  |  |
| --- | --- | --- |
|  | Miljoona lukuoperaatiota | Miljoona kirjoitusoperaatiota |
| Managed Disks/ korkealuokkainen SSD | 0€ | 0€ |
| Managed Disks/ SSD | 0.19€ | 0.19€ |
| Managed Disks/ HDD | 0€ | 0€ |
| Files/ korkealuokkainen | 0€ | 0€ |
| Files/ tapahtuma optimoitu | 0.14€ | 1.39€ |
| Files/ viileä | 1.21€ | 12.02€ |
| Blob Storage/ korkealuokkainen | 0.17€ | 2.11€ |
| Blob Storage/ Kuuma | 0.47€ | 6.01€ |
| Blob Storage/ arkisto | 600.0€ | 12.02€ |
| Cosmos | 0.23€ | ~0.50€ |
| SQL | ~0.34€ | ~0.70€ |

Cosmos tietokannan lukuoperaatioiden hinnan voi laskea suoraan, koska jokainen lukuoperaatio käyttää aina vain yhden RU-resurssin. Kirjoitus- operaatio voi käyttää eri määrän RU-resursseja riippuen tietokannan koosta ja kirjoitettavan tiedoston koosta. Taulukossa on arvioita että kirjoitus-operaatio käyttäisi keskimäärin 2 RU-resurssia. SQL- tietokannan IO-operaatoiden hinta on vieläkin vaikemapi arvioida, koska vastaava laskennan abstraktimitta DTU-yksikkö laskutetaan eri tavalla. DTU- yksiköistä laskutetaan käytettävistä yksiköistä kuukaudessa. Taulukun hinta on arvioitu 5 DTU-yksikön tietokannalla arvioiden että yksi lukuoperaatio vie yhden yksikön ja operaatio käyttää yksikköä sekunnin ajan. Näin viiden yksikön tietokannalla pystyttäisiin suorittamaan noin 13 miljoonaa lukuoperaatiota kuukaudessa. Kuukauden hinta on jaettu 13, josta on saatu miljoonan lukuoperaation hinta. Kirjoitus operaatioiden on arvioitu käyttävän resursseja tuplasti ja näin ollaan päädytty taulukon hintaan. Nämä hinnat edustavat minimihintaa, ja edellyttyvät että tietokannasta käytettäisiin 100% teholla. Todelliset hinnat tulisivat näin ollen olemaan kalliimpia, riippuen tietokannan käyttöasteesta.

Kun vertailee taulukkoa 1 ja taulukkoa 2 niin huomaa selkeästi logiikan kylmien ja kuumien talletusratkaisuiden välillä. Kuumat ratkaisut ovat IO-suorituskyvyltään tehokkaita ja talletustila kustantaa enemmän, mutta IO-operaatiot ovat halpoja. Kylmät ratkaisut ovat päin vastoin erittäin halpoja talletustilaltaan, mutta IO-operaatiot ovat erittäin kalliita ja latenssi erittäin suuri. Kun halutaan optimoida talletusratkaisun hintaa, täytyy miettiä useampia asioita, jotka vaikuttaa hintaa. Kuinka paljon IO-operaatioita suoritetaan ja onko pieni latenssi ja hyvä suorituskyky kriittinen ominaisuus. Jos IO-operaatioita on paljon, mutta nopeus ei ole kriittinen tekiä, optimaalisin hintaratkaisu löytyy lämpimän päädyn ratkaisuista. Jos IO-operaatioita ei ole paljon ja nopeus ei ole kriittinen tekijä, kannattaa valita kylmämalli. Taulukoista erottuu selvimmin Blob Storagen arkisto- malli. Sen talletustilakustannukset ovat selvästi muita ratkaisuja halvemmat, mutta lukuoperaatiot ovat puolestaan luokkaa tuhat kertaa kalliimmat kuin minkään muun ratkaisun. Tietokantojen(Cosmos ja SQL) kustannukset ovat kalliimmat kuumilla ratkaisuilla, mutta ne mahdollistavat monimutkaisten kyselyjen tekemisen ja rinnakkaisen käytön turvallisesti.

### Tiedostojen siirron kustannukset

Tiedostojen siirron kustannuksia vertaillaan epäsuorasti: kuinka paljon eri siirtoprotokollat kasvattavat tiedoston kokoa siirrettäessä. Siirrettävän datan kasvu aiheuttaa lisää kustannuksia verkkoliikenteen kasvaessa. Pilvialustat velottavat virtuaalikoneen tulevasta ja lähtevästä datasta.

Tutkimuksessa käytettiin WireShark nimistä ohjelmistoa. Se on avoimen lähdekoodin pakettianalysaattori, jonka avulla voidaan tutkia verkkoliikennettä. Wireshark on monialustainen ohjelmisto ja se käyttää PCAP:ia (Packet Capture) pakettien sieppaamiseen. PCAP on verkkokäytäntö, joka kopioi verkossa kulkevia datapaketteja niiden liikkuessa verkon poikki ja tallentaa syntyneet kopiot analysointia varten. Wireshark tarjoaa käyttöliittymän kaapattujen pakettien analysointiin, jonka avulla projektin tietoliikennettä on tutkittu.

Projektissa tutkittiin tietoliikennettä tiedostoa siirtäessä molempiin suuntiin: palvelimelle ja palvelimelta kulkevaa liikennettä. Molempiin suuntiin kulkenut data laskettiin yhteen ja vertailu tehtiin tällä summatulla arvolla. Vertailussa on mukana kaikki projektissa käytetyt protokollat ja niitä on käytetty C# -kirjastoilla. Vertailu tehtiin kahden erikokoisella tiedostolla. Tiedostojen koot olivat 1M tavua sekä 200M tavua. Kuvassa X on esitetty vertailu 1 M tavun tiedostolla sekä kuvassa Y on esitetty vertaiulu 200M tiedostolla. Tiedoston kasvu on kuvissa ilmoitettu prosenttiaalisena kasvuna verrattuna alkuperäiseen tiedostoon.

A graph of different colored bars

Description automatically generated

A graph of different colored bars

Description automatically generated

1 M tiedosto kasvaa hieman enemmän siirtäessä kuin suurempi 200 M tiedosto kaikilla muilla protokollilla, paitsi FTP:llä jolla se pysyy melkein samana. 1M tiedoston kasvuprosentti on 3.7 - 6.7 % välillä ja 200 M tiedoston 4.0 - 4.7 % välillä. SFTP kasvaa eniten kummallakin tiedostokoolla. FTP kasvaa vähiten pienemmällä tiedostolla ja WebSocket puolestaan isommalla tiedostolla. Salatut protokollat kasvavat kaikista eniten isommolla tiedostolla ja pienemmällä tiedostolla SFTP ja HTTPS kasvavat eniten, mutta FTPS kasvaa vähemmän kuin salaamattomat HTTP tai WebSocket. Pienellä tiedostolla kasvuun vaikuttaa enemmän esimerkiksi protokollan yhteydenavaus ja käyttäjänhallinan prosessit. Esimerkiksi HTTP protokollaa kasvattaa HTTP -pyynnön vastaus. Pienemmällä tiedostokoolla tämä vastaus on noin 1% tiedoston koosta.

[https://www.forbes.com/advisor/business/software/what-is-pcap/]

## Tehokkuusvertailu

Kuva 12 esittää FTP-, FTPS- sekä SFTP-latausten latausaikojen tulokset Python- ja C#-kirjastoilla toteutettuna. Kuvasta nähdään että SFTP on selvästi FTPtä ja FTPStä hitaampi molemmilla ohjelmointikielillä suurimmilla tiedostoilla. Pienellä tiedostokoolla SFTP ei ole hitaampi kuin FTP tai FTPS, mutta ero kasvaa voimakkaasti kun tiedoston koko kasvaa. Python koodissa SFTP:n suorituskyky on selvästi huonompi, kuin mitä C#-koodilla on, mutta FTP- ja FTPS-latausten kohdalla ohjelmointiielellä ei ole merkitystä. FTP ja FTPS suorituskyvyt eroavat toisistaan vain vähän. Pienillä tiedostoilla FTP on hieman nopeampi koska aloitusprosessi on lyhyempi, koska yhteyttä ei salata.

Chart, line chart

Description automatically generated

Kuva 15 FTP, FTPS ja SFTP latausten tulokset

Kuva 13 esittää HTTP latauksien latausajat. Latausten pyynnön koko on esitetty x-akselilla ja latausaika on esitetty y-akselilla. Pylväiden väri puolestaan merkitsee eri pyynnön koodausta. Kuvaajasta näkee selvästi että pyynnön koon suurentuessa latausaika pienenee erittäin voimakkaasti kaikilla koodauksilla. Pienempi pyynnön koko aiheittaa useampia pyyntöjä palvelimelle, joka hidastaa latausta merkittävästi. 5 kB:n kokosilla pyynnöillä hitain koodaus on base64, vaikka teoriassa array on kaikista hitain koska se kasvattaa eniten siirrettävän datan kokoa.

Chart, bar chart

Description automatically generated

Kuva 16 HTTP-latausten latausajat

Kuva 14 esittä HTTP -latausten lähetettävän datan määrää suhteessa tiedoston kokoon. Lähetettävä data on mitattu Wireshark -ohjelmalla jota käytetään verkkojen analyysiin. Siirrettävässä datassa on mukana kaikki TCP- tasolla siirtnyt data, joten se katttaa siirretävän tiedosotn sekä HTTP- pyyntöön tarvittavan datan. Kun käytetään pientä 5kb kokoista pyyntöä lähetettävän datan määrä kasvaa reilusti kaikilla koodauksilla. Teoriassa form-data on kaikista tehokkain tapa lähettää dataa ja data siirtyy binäärimuodossa. Form-data kasvattaa lähetettävän datan määrää vain muutamalla prosentilla kun lähetettävän pyynnön koko on suuri. Base64 koodaus kasvattaa lähetettävän tiedoston kokoa noin 33% kun käytetään suurta pyyntöä. Array koodaus kasvattaa siirrettävän datan 100% suuremmaksi mitä alkuperäinen tiedosto, jonka takia se ehdottomasti huonoin tapa siirtää tiedostoja HTTP protokollan avulla. Ainoastaan 500 kB on nähtävissä selkeasti teoreettinen tulos että form-data on nopein jonka jälkeen base64 koodaus on noin 33% hitaampi sekä array useampi kertoja hitaampi.

Chart, bar chart

Description automatically generated

Kuva 17 HTTP -latausten tiedoston kasvu lähettäessä

Kuva 15 esittää Webocket-latauksen tulokset. Websocket-yhteys luotiin .NET:n SignalR-kirjaston avulla. SignalR-kirjasto on osa .NET joten sitä ei tarvitse erikseen asentaa palvelimelle, mutta selaimeen se tarvitsi asentaa npm-pakettina. Signal käyttää konepellin alla selaimen Websocket API:a, mutta se luo korkeamman tason rajapinnan, mahdollistaen palvelinpään funktioiden kutsumisen käyttöliittymästä websocket-yhteyden avulla. Binääriviestintää varten SignalR kirjastoon joutuu asentamaan lisäosan Nuget paketin nimeltään MessagePack sekä käyttöliittymään saman nimisen npm-paketin. Lataus suoritettiin sekä merkkijono viestinä että binäärinä ja testissä käytettiin kolmea erikokoista tiedostoa. Latausten ajat olivat hyvin samanlaiset riippumatta käytettiinkö merkkijono- vai binääriformaattia. Pienimmillä tiedostoilla merkkijono oli hitusen nopeampi, mutta suuremmalla tiedostolla binääri oli hieman nopeampi. Teoriassa binääriformaatti pitäisi olla nopeampi, koska siirrettävä data on pienempi.

Chart

Description automatically generated

Kuva 18 Websocket-lataus

**Error! Reference source not found.** esittää siirrettävän datan suuruutta verrattuna alkuperäiseen tiedostoon. Siirrettävä data on mitattu TCP -tasolla, samaan tapaan kuin HTTP -pyyntöjen siirrettävän datan suuruutta vertailtaessa. String -formaatilla lähetettäessä siirrettävä data on kaksinkertainen alkuperäiseen tiedsotoon nähden ja binary -formattissa siirrettävä data 11% suurempi mitä alkuperäinen tiedosto.

Chart

Description automatically generated

Kuva 19 Websocket -latauksen siirrettävän datan kasvu

Kuva 17 esittää tiedostojen talletusnopeuksia eri talletusratkaisuilla. Kuvan y-akseli on logaritminen ja siitä erottuu selvästi etttä Disk ja File Storage ovat huomattavasti nopeimpia talletusratkaisuja. Tähän vaikuttavia tekijöitä on sisäverkon verkkoliikenne ja ne eivät vaadi mitään kirjautumismekanismia. SQL-talletus on seuraavaksi nopein. Siinä data siirretään TCP yhteydellä sisäverkossa, mutta SQL-kyselyt tarvitsevat kirjautumis- ja datan eheysprosessit jotka hidastavat tiedostojen tallettamista. Blob Storage sekä Cosmos käyttävät julkisen verkon HTTP API:a tiedostojen tallentamiseen ja ovat siksi ehdottomasti hitaimpia talletuspaikkoja. Blob Storage mahdollistaa binääritiedostojen tallentamisen binäärimuodossa, mutta Cosmos käyttää pelkästään json-dokumenttejä ja tiedostojen koko kasvaa merkittävästi kun on tallennettuna tekstinä. Cosmoksen json-dokumentin maksimikoko on 2MB joten Cosmoksen suorituskykyä ei voitu testata 10 MB tiedostoilla.

Chart, line chart

Description automatically generated

Kuva 20 Tiedostojen talletusnopeuksia

Yksi tutkimuksista on vertailla tiedostojen tallentamista SQL-kantaan binäärisenä blob tyyppina. Toinen mahdollinen vaihtoehto on tallentaa SQL-tietokantaan tiedoston tiedostopolku ja tallentaa tiedosto kovalevylle tietokannan sijaan. Kuva 18 esittää tutkimuksen tulokset. Tutkimuksessa vertailtiin tiedostojen lukua, kirjoitusta, poistoa sekä tiedosto infon lukua tietokannasta. Tutkimuksessa käytettiin 2 tietokanta taulua jotka ovat muuten identtiset, mutta toisessa tiedosto on tallennettuna tietokantaan ja toisessa linkin avulla tiedostojärjestelmään. Tiedostoinfo on tallennettuna samalla rivillä tiedoston tai sen linkin kanssa ja tyypiltään se on varchar. Tuloksista nähdään selvästi että tiedoston luku selvästi nopempi koska erillistä levylukuoperaatiota ei tarvitse tehdä tietokanta kyselyn jälkeen. Infon luku on hieman nopeampi taulussa jossa on linkki, johtuen että rivien indeksointi on hieman nopeampaa RAM muistissa koska rivien koko on pienempi. Kirjoitus- ja poistonopeudet ovat lähes samat riippumatta tallennustavasta. Toinen huomioita asia tiedostojen tallentamisessa levyjärjestelmään on tietokannan eheys. Kun tiedostot tallennetaan tietokantaan niin tietokanta pysyy eheänä tiedostojen suhteen. Levylle tallennettaessa eheys saattaa rikkoontua koska tietokantaoperaatio ja levyoperaatio eivät ole saman tietokantatransaktion sisällä

Chart, bar chart

Description automatically generated

Kuva 21 Tiedostojen tallentaminen SQL-tietokantaan ja kovalevylle

## Tietoturvavertailu

### Suojaamattomat protokollat

Tietoturvavertailussa vertaillaan eri talletustapojen sekä tiedonsiirtoprotokollien tietoturvallisuutta. Käsitellyistä tiedonsiirtotavoista suojaamattomia protokollia ovat HTTP, FTP, WebSocket ja WebDav. Niiden käyttöä ei suositella nykyään olleenkaan, vaan niiden tilalla pitäisi käyttää suojattuja versioita HTTPS, FTPS WebsocketS ja WebDavS. Kaikki suojaamattomat protokollat eivät salaa tiedonsiirtoa palvelimen ja asiakkaan välillä tämän takia kuka tahansa joka pääsee verkkoliikenteeseen käsiksi voi siepata ja lukea sen. Lisäksi suojaamattomat protokollat eivät tarjoa mitään todennusta, mikä tekee niistä haavoittuvia välimieshyökkäyksille (MITM). MITM -hyökkäys on yleinen termi, jossa tekijä päsee kuuntelemaan ja muokkaamaan kahden käyttäjän välistä keskustelua. Tarkoituksena on salakuunnella keskustelua tai esiintyä toisena osapuolena muokkaamalla keskustelua. jolloin se vaikuttaa tavalliselta tiedonvaihdolta hyökkäyksen kohteeksi joutuneille ja ne uskovat kommunikoivan suoraan keskenään, vaikka hyökkääjä on asettunut kahden käyttäjän väliin.

HTTP- protokollassa on lisäksi haavoittuvuuksia, joita ei pystytä torjumaan TLS- salauksella. BREACH haavoittuvuus johtuu HTTPP- protokollan pakkausalgoritmista. HTTP-pakkaus suoritetaan tavallisesti deflate-algoritmin avulla. Kun dataa pakataan tällä algoritmilla, kaikki toistuvat tavusekvenssit havaitaan, eikä niitä toisteta pakatussa datassa. Sen sijaan toistuva tavusekvenssi tallennetaan vain kerran, samoin kuin osoittimet, jotka osoittavat mistä sama sekvenssi löytyy uudelleen. Tämä vähentää lähetettävien tavujen määrää. Pakatun datan pituus on kuitenkin näkyvissä myös salattuna, ja tämä on yksi peruselementeistä, joka tekee BREACH-hyökkäyksen mahdolliseksi. Jotta sovellus olisi haavoittuvainen BREACH-hyökkäykselle, palvelimen HTTP-pakkaus täytyy olla kytkettynä päälle. hyökkääjällä tulee olla keino tarkastella uhrin liikennettä ja lisäksi hyökkääjällä on oltava mahdollisuus lähettää HTTP-pyyntöjä haavoittuvaan palvelimeen. Hyökkäys toimii lähettämällä pyyntöjä palvelimelle saadakseen tietoja pakatusta ja salatusta vastauksesta. Vastauksen otsikoita ei pakata, joten tällä hyökkäyksellä niitä ei pystytä kaappaamaan. Esimerkiksi evästeet siirretään otsikoissa, mutta CSRF-avaimet ovat saatavilla tällä hyökkäyksellä. Esimerkiksi CSRF-avaimen varastamiseksi hyökkääjän tulee tietää missä muodossa CSRF-avain lähetään HTTP- pyynnöissä ja vastauksissa. esimerkiksi voidaan käyttää CSRF avaimelle muotoa token=123456789. Hyökkääjä aloittaa ensimmäisen numeron arvaamisen ja lähettää palvelimelle pyyntöjä joissa on pelkästään mukana arvattavan CSRF- avaimen ensimmäinen numero. Hyökkääjä lähettää kaikki mahdolliset numerot palvelimelle. Palvelimen vastauksista lyhin on se missä numero on arvattu oikein. Tämä prosessi toistetaan jokaiselle numerolle erikseen kunnes koko CSRF- avain on saatu muodostettua. Hyökkäykseltä pystyy suojautumaan laittamalla palvelimen HTTP-pakkauksen pois päältä. <https://www.acunetix.com/blog/articles/breach-attack/>

### TLS

Kaikkien protollien suojaamiseen käytetään samaa TLS-salausprotokollaa, joten ne ovat kaikki haavoittuvia TLS-protokollan tietoturvaongelmille. Vanhoissa SSL/TLS versioissa on useita tiedossa olevia haavoittuvuuksia alle on esitelty muutama yleisesti tiedossa oleva hyökkäys protokollia vastaan.

RC4 hyökkäys symmetristä salausta vastaan. RC4 on symmetrinen virtasalausalgoritmi, joka julkistettiin vuonna 1987. RC4 on ollut mahdollinen symmetrinen salaus jo ensimmäisessä SSL 2.0 versiossa ja se poistettiin TLS 1.3 versiosta. Toisin kuin nykyaikaisessa virtasalauksessa, RC4 ei ota erillistä satunnaista alustusvektoria avaimen rinnalla. Tämä tarkoittaa, että jos yhtä pitkäaikaista avainta käytetään useiden virtojen turvalliseen salaamiseen, alustusvektorin ja pitkän aikavälin avain on yhdistetään RC4:n salausavaimen luomiseksi. Ensimmäiset haavoittuvuudet RC4 algoritmista löydettiin vuonna 2001 ja 2005. Haavoittuvuudet mahdollistivat löytämään käytetyn salausavaimen suuresta määrästä samalla avaimella salattuja viestejä, jos avain ja alustusvektori oli vain liitetty toisiinsa salausavaimen luomisessa. Huolimatta RC4:n tietoturvaa rikkovista hyökkäyksistä, RC4:ään perustuvia SSL- ja TLS-salausohjelmistoja pidettiin edelleen turvallisina ennen vuotta 2013 sen perusteella, miten niitä käytettiin SSL:ssä ja TLS:ssä. Vuonna 2011 RC4 suositeltiin BEAST-hyökkäyksen kiertäjäksi. Vuonna 2013 julkistetut uudet hyökkäysmuodot osoittivat lopullisesti TLS:n RC4:n rikkomisen toteutettavuuden. Hyökkäys vaatii 2^34 salattua viestiä RC4:n murtamiseen ​​TLS:ssä ja SSL:ssä. Hyökkäyksestä julkaistiin vuonna 2015 uusi versio, joka vaatii enää 2^26 viestiä salauksen murtamiseen. <https://beaglesecurity.com/blog/vulnerability/the-rc4-algorithm-in-transport-layer-security-and-secure-sockets-layer.html>

BEAST on selainhaavoittuvuus, joka julkaistiin vuonna 2011 ja se koskee SSL 3.0:aa ja TLS 1.0:aa. Hyökkääjä voi purkaa kahden osapuolen välillä vaihdetun tiedon salauksen hyödyntämällä haavoittuvuutta lohkosalauksen ketjuttamisen CBC (Cipher Block Chaining) toteutuksessa. Tämä on asiakaspuolen hyökkäys, joka on välimieshyökkäys. Hyökkääjä lisää paketteja TLS-virtaan. Tämän avulla he voivat arvata muokatun viestin kanssa käytetyn alustusvektorin ja siten purkaa salauksen. Jotta BEAST-hyökkäys onnistuisi, hyökkääjällä on oltava jonkinlainen hallinta uhrin selaimessa. Selainten modernit versiot estävät BEAST- hyökkäyksen. RC4 virtasalaus on immuuni BEAST-hyökkäykselle. Siksi RC4:ää käytettiin laajalti keinona lieventää BEAST-hyökkäystä palvelinpuolella ennen kuin se todettiin murretuksi vuonna 2013. <https://www.acunetix.com/blog/articles/tls-vulnerabilities-attacks-final-part/>

POODLE on SSL 3.0 haavoittuvuus, joka johtuu BEAST:in tavoin lohkosalauksen ketjuttamisen CBC (Cipher Block Chaining) toteutuksesta. POODLE on välimieshyökkäys ja edellyttää, että hyökkääjän JavaScript -koodi ajetaan uhrin selaimessa. POODLE haavoittuvuus johtuu lohkosalausalgoritmien täytetarpeesta. Salattavan datan pitää olla lohkosalausalgoritmeissa lohkokoon monikerta, muuten viimeinen lohko pitää täyttää täytteellä loppuun. Jos data on täsmälleen lohkokoon monikerta, silloin viimeinen lähetettävä lohko on pelkästään täytettä. Palvelin hylkää pyynnön jos täyte on viallinen pyynnössä. POODLE hyökkäys perustuu siihen että hyökkääjä pääsee muokkaamaan lähetettävän pyynnön kokoa ja tietää siten pyynnön täytteen. Täytettävä sisältävä lohko on viimeisenä ja hyökkääjä vaihtaa viimeisen lohkon ja murrettavan lohkon paikkaa ja lähettää pyynnön useaan kertaan palvelimelle muuttaen pyynnön kokoa jokaisella pyynnöllä. Jos palvelin hyväksyy pyynnön, niin murrettavan lohkon bitit vastaavat täytelohkon täytettä vastaavia bittejä. Hyödyntämällä tätä haavoittuvuutta hyökkääjä saa murrettua SSL:n salauksesta lohkoja. Käyttäjän evästeiden varastaminen on yksi esimerkki onnistuneesta POODLE- hyökkäyksestä. Esimerkiksi 16 bittisen evästeen varastamiseen tarvitaan maksimissaan 4096 pyyntöä, jos hyökkääjä tietää tarkalleen missä kohtaa pyyntöä evästeet lähetetään. <https://www.acunetix.com/blog/web-security-zone/what-is-poodle-attack/> <https://www.acunetix.com/blog/articles/poodle-gives-final-bite-puts-sslv3-rest/>

DROWN on SSL 2.0 version haavoittuvuus, joka johtuu protokollan liian heikosta symmetrisestä salauksesta. Hyökkääjän on ensiksi kuunneltava useita satoja SSL-yhteyksiä uhrin ja palvelimen välillä. Yhteyksien tarvitsee jakaa jaettusalaus RSA-kättelyllä. Hyökkääjä lähettää palvelimelle muokattuja versioita uhrin salatuista RSA-kättelyistä. Tapa jolla palvelin vastaa kuhunkin näistä muokatuista kättelyistä riippuu siitä muodostuuko salatusta viestistä oikean muotoinen viesti salauksen purkamisen jälkeen. Salaus käyttää vain 40 bittistä avainta RSA-salauksessa joten hyökkääjä pystyy vertaamaan palvelimen vastausta kaikkiin mahdollisiin vastauksiin joita on 2^40 kappaletta. Tällä tavalla hyökkääjä pystyy päätteleen oliko muokattu versio salatusta kättelystä oikein muodostettu vai ei. <https://drownattack.com/>

Tiedossa olevat TLS-haavoittuvuudet koskevat pääasiassa vanhentuneita TLS-versioita, joiden käyttämistä ei enää suositella. RC4- salauksen haavoittuvuus koskee myös turvallista TLS 1.2 versiota. Kuva 20 esittää Acunetix Web Application Vulnerability Report- tutkimuksen tuloksia vuodelta 2020. Tutkimuksen on tehnyt Invicti niminen tietoturvayhtiö ja siinä on testattu 5000 palvelimen tietoturvaa. Tutkimukseksen perusteella palvelimet sallivat edelleen TLS-yhteyden muodostamisen protokollan vanhoilla versioilla. POODLE- hyökkäys edellytttää SSL 3.0 versiota ja DROWN puolestaan vanhempaa SSL 2.0 versiota. Silti palvelimista 3.9% oli haavoittuvia POODLE- hyökkäykselle ja 0.7% oli haavoittuvia DROWN- hyökkäykselle. BREACH- hyökkäys johtuu HTTP- protokollan pakkausalgoritmista ja sille oli haavoittuvia 3.9% palvelimista. 7.7% palvelimista salli symmetriseksi salaukseksi käyttää murrettua RC4- algoritmia. RC4- algoritmia on mahdollista käyttää kaikissa muissa versioissa paitsi TLS 1.3 versiossa. Sen käyttö pitäisi erikseen kieltää TLS- palvelimen asetustiedostossa, jotta haavoittuvuudelta vältyttäisiin. Lisäksi 30.7 % palvelimista salli TLS 1.0 version käyttämisen joka on haavoittuvainen muun muassa BEAST- hyökkäykselle.

https://www.acunetix.com/white-papers/acunetix-web-application-vulnerability-report-2020/

A pie chart with different colored sections

Description automatically generated

Kuva 22 Vuoden 2020 AWAVR- tutkimuksessa löydetyt TLS/SSL- haavoittuvuudet

Kuva 21 esittää F5 Labsin tekemän The 2021 TLS Telemetry Report- tutkimuksen tuloksia HTTPS- palvelimien tukemista SSL/TLS- versioista. Tutkimuksessa on käytetty miljoona suosittua HTTPS-sivustoa. Tutkimuksesta huomataan että SSL 3.0- version käyttämisen mahdollistaa noin muutama prosentti tutkituista palvelimista. Nämä palvelimet ovat todennäköisesti haavoittuvia POODLE- hyökkäykselle. TLS 1.0 tuki on noin 50% palvelimista, mikä on huomattavasti suurempi luku mitä Acunetix Web Application Vulnerability Report- tutkimukseksessa saatu 30.9% on. TLS 1.1 on muutaman prosentin tuetumpi versio mitä TLS 1.0- versio. Ehdottomasti tuetuin versio on TLS 1.2, joka saavuttaa melkein 100% tuen. Uusin versio 1.3 saavutti noin 60% tuen tutkimuksessa. Vaikka F5 Labsin tekemä tutkimus ei suoraan paljasta tietoturvahaavoittuvuuksia TLS- prokollassa, niin se antaa hyvän kuvan kuinka laajasti protokollan vanhentuneet tietoturvaongelmia sisältäviä versioita käytetään jotka mahdollistavat TLS-protokollaan kohdistettuja hyökkäyksiä. https://www.f5.com/labs/articles/threat-intelligence/the-2021-tls-telemetry-report

A graph of different colored lines

Description automatically generated

Kuva 23 Suosituimpien HTTPS- palvelimien tukemat TLS- versiot vuonna 2021

Kun vertaillaan FTPS- ja SFTP-protokollia tietoturvan näkökulmasta, oleellisen asia on käytettävien porttien määrä. SFTP käyttää vain yhtä porttia tiedonsiirtoon, mutta FTPS tarvitsee kaksi. Lisäksi jokainen rinnakkainen yhteys tarvitsee oman porttinsa. Tämän takia FTPS -palvelimelle määritellään usein sallittujen porttien joukko palomuuriin. Palomuurit voidaan katsoa putkena, jotka määrittelevät sallitun ja kielletyn verkkotoiminnan yksityisessä verkossa. Palomuurit luovat tarkastuspisteitä verkkoliikenteen ohjaamiseksi, jolloin ne tarkistavat ohjelmoitujen parametrien perusteella verkkoliikennettä, ja toimivat niiden mukaisesti. Mitä vähemmän portteja on avattu, sitä vähemmän järjestelmässä on haavoittuvia kohtia. Jos käytössä on palomuuri, niin se pitää konffiguroida hyväksymään kaikki FTPS-palvelimen käyttämät portit, mikä lisää uhkia. Lisäksi salaus kiinnitetään FTPS -protokollaan TLS- tai SSL-tekniikoilla FTP -protokollan päälle. Tämän ulkoisen salauksen seurauksena palomuuriratkaisut eivät pysty havaitsemaan, mitä porttia käytetään ja miksi. Yksiporttiset SFTP-asetukset ovat ihanteellisia käytettäväksi palomuurin rinnalla. Se muodostaa yhden yhdistetyn yhteyden asiakkaan ja palvelimen välille. Palomuuri voi tarkkailla tämän yhteyden poikkeavuuksia, epäilyttäviä merkkejä ja muita uhkamerkkejä. [https://www.spiceworks.com/tech/networking/articles/sftp-vs-ftps/]

### Microsoft Azure

Azure käyttää roolipohjaista pääsynhallintaa (Azure RBAC) Azure-resurssien käyttöoikeuksien hallintaan. Rooleja voidaan antaa käyttäjille, ryhmille, sovelluskäyttäjille tai hallituille identiteeteille. Käyttöoikeudet resursseihin on sidottu rooleihin. Azure tarjoaa valmiiksi määriteltyjä rooleja ja käyttäjät voivat luoda uusia itse määriteltyjä rooleja. Rooleille on määritettävä laajuus. Laajuus on joukko resursseja, joita käyttöoikeus koskee. Azuressa laajuuden voi määrittää neljällä tasolla laajasta kapeaan: hallintaryhmä, tilaus, resurssiryhmä ja resurssi. Kuva 22 esittää Azuren laajuuksien hierarkian.

A diagram of a group

Description automatically generated

Kuva 24 Azuren roolien laajuus

Alimpana hierarkiassa on yksittäinen resurssi, joka tarkoittaa jotakin Azuren tarjoamaa palvelua. Sen yläpuolella on resurssiryhmä, jonka avulla resursseja voidaan ryhmitellä. Jokaisen resurssin on kuuluttava yhteen ryhmään ja ryhmässä voi olla useampia resursseja. Resurssiryhmän yläpuolella on tilaus, johon voi kuulua useampi resurssiryhmä. Kun käyttäjätili luodaan Azureen, käyttäjätili sisältää yhden tilauksen. Tilauksen sisältämistä resursseista muodostetaan lasku ja statistiikkaa kuluista ja käytettyistä palveluista. Käyttäjä tiliin voidaan luoda lisää tilejä, joille kaikille muodostuu oma laskunsa. Useat tilit voidaan vielä yhdistää toisiinsa hallintaryhmän avulla. <https://learn.microsoft.com/en-us/azure/role-based-access-control/scope-overview> <https://learn.microsoft.com/en-us/azure/role-based-access-control/role-assignments-portal> <https://learn.microsoft.com/en-us/azure/role-based-access-control/role-assignments-steps>

Kaikki käyttäjät, joilla on käyttöoikeus resurssiin voi käyttää resurssia omien Microsoft läyttäjätunnusten avulla, joko selaimessa toimivan Azure Portalin avulla tai CLI- työkalun kautta. Molemaat työkalut tarvitsevat käyttäjän käyttäjänimen ja salasanan. Kehitettävissä sovelluksissa käyttäjänimen ja salasanan käyttäminen ei ole suotavaa, koska oman salasanan suojaaminen muilta kehittäjiltä tai ylläpitäjiltä on erittäin haastavaa. Lisäksi käyttäjällä on yleensä laajemmat käyttöoikeudet resursseihin mitä kehitettävä sovellus vaatisi. Tätä käyttötapausta varten pääsynhallinnassa on sovelluskäyttäjät ja hallitut identiteetit. Sovelluskäyttäjät liittyvät Microsoftin laajempaan pääsynhallintajärjestelmään Microsoft Entra ID:hen. Microsoft Entra ID kattaa muitakin Microsoftin palveluita, kuten Office 365 ja OneDrive, joihin sen avulla voi antaa pääsyoikeudet. Sillä on oma selainpohjainen hallintapaneeli, jonka avulla voidaan antaa pääsyoikeuksia eri Microsoftin palveluihin. Sovelluskäyttäjille voidaan antaa Azuren roolien avulla käyttöoikeudet eri Azururen palveluihin. Hallitut identiteetit ovat sovelluskäyttäjien erikoistapauksia ja ne toimivat samalla tavalla. Erona on vain että hallitut identiteetit ovat olemassa vain Azuren sisällä ja niillä voidaan antaa käyttöoikeusia vain Azuren resursseihin. Sovelluskäyttäjät ja hallitut identiteetit mahdollistavat kirjautumisen sertifikaatin tai pitkän satunnaisgeneroidun salasanan avulla. Niille voidaan antaa juuri sovelluksen tarvittavat käyttöoikeudet ja käyttöoikeusia voidaaan muokata aina tarvittaessa. <https://learn.microsoft.com/en-us/entra/identity/managed-identities-azure-resources/overview> <https://learn.microsoft.com/en-us/azure/role-based-access-control/scope-overview> <https://learn.microsoft.com/en-us/entra/identity-platform/howto-create-service-principal-portal> <https://www.microsoft.com/en-us/security/business/identity-access/microsoft-entra-id>

Azure tarjoaa mahdollisuutta hallita yksittäisen talletusratkaisun pääsynhallintaa 512- bittisen käyttöavaimen avulla. Käyttöavaimia on kaksi, jotka mahdollistavat sekä kirjoitus että lukuoperaatiot sekä kaksi avainta jotka mahdollistavat pelkät lukuoperaatiot. Käyttöavaimet ovat käytössä seuraavissa talletusratkaisuissa: Cosmos, File Share ja Blob Storage. Käyttöavaimen avulla talletusratkaisua voidaan käyttää sen julkisen päätepisteen tai ohjelmistokielien SDK- työkalujen avulla molemmat lukuoikeuden antavat avaimet mahdollistavat samallaiset lukuoidet talletusratkaisuun. Samalla tavalla kaksi kirjoitus ja lukuoikeutta antavaa avainta antaa samat oikeudet. Kaksi kumpaakin avainta mahdollistaa avainten kierrättämisen. Kierrättämisellä tarkoitetaan toisen avaimen päivittämistä uuteen avaimeen. Jos sovelluksella on käytössä molemmat avaimet, niin se pystyy toimimaan normaalisti käyttämällä toista vanhempaa vielä voimassa olevaa avainta. Kierrättämisellä pyritään parantamaan tietoturvaa ja Microsoft suosittelee säännöllistä avainten kierrättämistä. Avainten avulla voidaan luoda jaettuja käyttöoikeusallekirjoituksia (SAS). Jaettu käyttöoikeusallekirjoitus on URI, joka myöntää rajoitetut käyttöoikeudet talletusratkaisuun. Jaettuun käyttöoikeusallekirjoitukseen voidaan määrittää mitä operaatioita sen avulla voidaan suorittaa sekä voimassaoloaika. Jaetun käyttöoikeusallekirjoitus sopii tilanteihin joissa halutaan antaa pääsy vain tiettyihin tallennustilien resursseihin tietyn ajan. <https://learn.microsoft.com/fi-fi/azure/storage/common/storage-account-keys-manage?tabs=azure-portal#regenerate-access-keys> <https://learn.microsoft.com/en-us/azure/cosmos-db/secure-access-to-data?tabs=using-primary-key> <https://learn.microsoft.com/fi-fi/azure/storage/common/storage-account-keys-manage?tabs=azure-portal>

Microsoft Entra ID:n aikaisempi nimi oli Azure Active Directory (ADD) ja siitä löydettiin vakava haavoittuvuus vuonna 2023. Wiz- tutkimustiimi löysi haavoittuvuuden Microsoftin omistamasta Bing- hakukoneesta. Bing- hakukoneen pääsynhallinta oli toteutettu ADD -pääsynhallinnalla. Bing hakukoneen pääsynhallinnassa oli haavoittuvuus, joka mahdollisti minkä tahansa ADD -käyttäjän kirjautumisen Bing Trivia -sivustolle. Bing Trivia -sivusto oli hallintapaneeli, jonka avulla pystyttiin muun muassa muokkaamaan Bing hakukoneen hakusivun ulkoasua. Wiz- tutkimustiimi onnistui lisäämään Bing-hakukoneeseen JavaScript koodia. Hyökkäystä kutsutaan Cross Site Scripting (XSS), kun hyökkääjä onnistuu samaan haitallista JavaScript koodia ajoon uhrin selaimeen. XSS- hyökkäyksellä onnistuttiin varastamaan Bing käyttäjien Office 365 -tileihin kirjautumistunnuksia. Myös muihin pienempiin Microsoftin sisäisiin palveluihin pääsi kirjautumaan millä tahansa ADD- tunnuksella. https://www.wiz.io/blog/azure-active-directory-bing-misconfiguration

Azuren talletusratkaisuista on löytynyt myös vakava tietoturvahaavoittuvuus vuonna 2021, joka sai nimekseen ChaosDB. Haavoittuvuus koskee Cosmos tietokantaa ja sen löysi Wiz- tutkimustiimi. ChaosDB- haavoittuvuus mahdollisti rajoittamattoman pääsyn useiden tuhansien Azure -asiakkaiden Cosmos tietokantoihin. Haavoittuvuus johtui Jupiter Notebook -ominaisuuden virheellisestä konfiguraatiosta Cosmos tietokantaan. Jupyter Notebook on interaktiivinen verkkosovellus laskennallisten asiakirjojen luomiseen ja jakamiseen. Jupiter Notebook- ominaisuus mahdollistaa Cosmos tietokannan datan visualisoinnin. Hyödyntämällä Cosmos tietokannan Jupyter Notebook -ominaisuuden haavoittuvuuksien ketjua, hyökkääjä voi tiedustella tietoja kohde Cosmos tietokannan Jupyter Notebookista. Näin tehdessään hyökkääjä saa joukon tunnistetietoja, jotka liittyvät kohde Cosmos tietokannan tiliin sekä Jupyter Notebook -laskentaan. Näiden tunnistetietojen joukossa on myös Cosmos tietokannan käyttöavaimet, joiden avulla hyökkääjä pääsee käsiksi Cosmos tietokantaan. ChaosDB haavoittuvuus mahdollisti hyökkääjien nostamaan käyttöoikeuksiaan joka mahdollisti pääsyn Cosmos tietokannan sisäisiin salaisuuksiin ja varmentetihin. Näiden salaisuuksien avulla hyökkääjät pääsivät käsiksi muiden Cosmos tietokantojen käyttöavaimiin. Haavoittuvuuus koski kaikkia Cosmos tietokantoja, joissa oli Jupyter Notebook -ominaisuus päällekytkettynä. Jupyter Notebook -ominaisuus oli päällekytkettynä uusissa Cosmos tietokannoissa, ja käyttäjän piti erikseen laittaa ominaisuus pois päältä halutessaan tietokantaa luotaessa. Wiz- tutkimustiimi ilmoitti haavoittuvuudesta Microsoftille ja Microsodt korjasi ChaosDB -haavoittuvuuden kahden päivän kuluessa. Microsoft on ilmoittanut että sillä ei ole viitteitä että jokin muu taho kuin Wiz- tutkimustiimi olisi päässyt ChaosDB- haavoittuvuuden ansiosta käsiksi muiden käyttäjien Cosmos tietokantoihin. Microsoft kuitenkin suositteli kaikkia asiakkaitansa kierrättämään Cosmos tietokannan käyttöavaimensa. <https://securityaffairs.com/124510/hacking/chaosdb-flaw-technical-details.html> <https://chaosdb.wiz.io/>

Azure tarjoaa osalle talletusratkaisuista lisäpalveluna erillistä tietoturvapalvelua. Microsoft Defender for Storage on Azuren tietoturvakerros, joka havaitsee mahdolliset uhat talletusratkaisuissa. Se auttaa estämään seuraavia vakavia tietoturvaongelmia: haitalliset tiedostojen lataukset, arkaluonteisten tietojen vuotaminen ja tietojen korruptio. Sitä on mahdollista käyttää Blob Storage-, Azure Files- ja Azure Data Lake Storage – talletusratkaisuiden kanssa. Microsoft Defender for Storage sisältää seuraavat toiminnalisuudet: toiminnan seurannan, arkaluonteisten tietojen uhkien havaitsemisen ja Haittaohjelmien tarkistuksen. Toiminnan seurantaan Defender for Storage analysoi jatkuvasti suojattujen talletusratkaisuiden dataa ja ohjaustason lokeja, kun se on käytössä. Palvelu käyttää Microsoft Threat Intelligenceä tunnistaaksesi epäilyttävät asiat, kuten haitalliset IP-osoitteet, Tor-poistumissolmut ja mahdollisesti vaaralliset sovellukset. Se myös rakentaa tietomalleja ja käyttää tilastollisia- ja koneoppimismenetelmiä havaitakseen perustoiminnan poikkeavuuksia, jotka voivat viitata haitalliseen toimintaan. Palvelu luo tietoturvavaroituksia epäilyttävistä toiminnoista. Haittaohjelmien tarkistus Defender for Storagessa auttaa suojaamaan tallennustilejä haitalliselta sisällöltä suorittamalla täyden haittaohjelmatarkistuksen ladatulle sisällölle lähes reaaliajassa käyttämällä Microsoft Defender Antivirus -ominaisuuksia. Arkaluonteisten tietojen uhkien havaitseminen mahdollistapalvelun tuottamien tietoturvavaroituksien priorisoinnin tietojen arkaluonteisuuden avulla. Defender for Storagen hinta on 10€ kuukaudessa jokaista talletusratkaisun tiliä kohden. <https://learn.microsoft.com/en-us/azure/defender-for-cloud/defender-for-storage-introduction>

### Pilvialustojen turvallisuustutkimukset

Vaikka pilvialustojen palveluiden tietoturvaongelmat ovat erittäin harvinaisia niin pilvialustoille kehitettyjen sovellusten tietoturvaongelmat ovat yleisiä. Snyk tietoturvayhtiön vuonna 2022 tekemässä tutkimuksessa The State of Cloud Security Report 2022 tutkittiin pilvialustoille kehitettyjen sovellusten tietoturvahaasteita. Tutkimukseen osallistui tietoturva-asiantuntijoita sekä pilvikehittäjiä. Kuva 23 esittää tutkimukseen osallistuneiden koetut vakavat tietoturvaongelmat pilvialustoille kehitetyissa sovelluksissa. Yleisin ongelma oli järjestelmän seisokkki, joka oli johtunut virheellisestä konfiguroinnista. Tietomurrot ja tietovuodot olivat myös erittäin yleisiä, vastaajista noin 26% oli kokonut tietovuodon ja 33% oli kokenut tietomurron.

A graph with blue and white text

Description automatically generated

Kuva 25 Snykin tutkimukseen vastanneiden kokemat pilviturvallisuusongelmat

Samassa tutkimuksessa myös kysyttiin tietoturvaongelmien aiheuttajaa. Vastaajat saivat valita useampia teknisiä vaihtoehtoja, jotka olivat aiheuttaneet pilvialustoilla tietoturvaongelmia. Kuva 24 esittää tutkimuksen tuloksia teknisistä ongelmista, jotka olivat aiheuttaneet tietoturvahäiriöitä. Tutkimuksen mukaan eniten tietoturvaongelmia aiheutti turvattomasti säilytetyt varmuuskopiot, jotka olivat olleet mukana 23 % tapahtuneista tietoturvaongelmista. Ongelman aiheuttajat voidaan ryhmittää muutamaan pääryhmään. Tietoturvakriittisten asioiden turvaton säilyttäminen kuten varmuuskopiot ja API avaimet. Puutteelliset valvontamentelmät, esimerkiksi puutteelliset lokitiedot. Pääsynhallinnan virheelliset asetukset, sisältäen ryhmät sekä Object Storagen. Turvattomat yhteydet sekä tietovarastot(virtuaalikoneiden yhteydet, tiedonsiirto ja tietovarastot).

A graph with blue and white text

Description automatically generated

Kuva 26 Tietoturvaongelmien aiheuttajat Snykin tutkimuksessa

Samassa tutkimuksessa myös kysyttiin syitä, jotka olivat aiheuttaneet tietoturvaongelmia. Kuva 25 esittää syitä, jotka ovat aiheuttaneet tietoturvaongelmia kutkimuksen mukaan.

A graph with blue and white text

Description automatically generated

Kuva 27 Tietoturvaongelmien aiheutumisien syyt Snykin tutkimuksen mukaan

Tietämyksen puute tietoturvasta ja nopea ympäristön muuttuminen oli yleisimmät syyt, jotka oli johtanut tietoturvaongelmiin. Molemmat syyt todennäköisesti lisääntyvät tutkimuksen mukaan tulevaisuudessa kun yhä useammat tiimit ottavat käyttöön jatkuvan käyttöönottomenetelmän ja sen tuomat tietoturvahaasteet. Suurin osa syistä johtuu pelkästään ihmisten toiminnasta, eikä liity käytettyyn pilviteknologiaan. Inhimillisen tekijän vaikutusta tietoturvaongelmiin on tukittu IBM:n vuonna 2024 julkaisemassa tutkimuksessa 2024 Data Breach Investigations Report. IBM:n tutkimus käsittelee vuosina 2022-2023 tapahtuneita tietomurtoja ja se käsittää myös pilvialostojen ulkopuolisia tietomurtoja. Kuva 26 esittää IBM:n tutkimuksen tietomurtojen avaintekijöitä. Inhimillinen tekijä oli yleisin avaintekijä ja se oli mukana 68% tehdyissä tietomurroissa. Muut tekijät tutkimuksessa oli teknisiä, mutta niiden yleisyys oli maksimissaan puolet inhimillisestä tekijästä.

<https://go.snyk.io/rs/677-THP-415/images/cloud-security-report-22.pdf>

A graph with blue bars

Description automatically generated with medium confidence

Kuva 28 Avaintekijät tietomurroissa IBM:n tekemässä tutkimuksessa

Yksi IBM:n tutkimuksen inhimillisistä tekijöistä on tietojenkalastelu. Tietojenkalastelu tarkoittaa yritystä varastaa arkaluonteisia tietoja, tyypillisesti käyttäjätunnuksia, salasanoja tai pankkitunnuksia. Tietojenkalastelu tehdään yleensä sähköpostin tai tekstiviestien avulla. Tutkimuksessa vuonna 2023 vastaajista 20% ilmoitti saaneensa tietojenkalasteluviestejä ja 11% vastaajista oli avannut viestin. Tietojenkalastelun määrä on kasvanut tutkimuksen mukaan tasaisesti vuodesta 2016 asti, jolloitutkimuksessa alettiin pitää ensimmäistä kertaa kirjaa tietojenkalasteluviesteistä. Kuva 27 esittää IBM:n tutkimuksessa käytettyjä menetelmiä tietomurroissa. Yleisin mentelmä on käyttäjätunnukset, joka pitää sisällään virheellisen pääsynhallinnan sekä turvattoman tavan säilyttää käyttäjätunnuksia. Tietojenkalastelu on tutkimuksen mukaan toisista yleisin tietomurtojen aiheuttaja haavoittuvuuksien hyväksikäytön kanssa.

<https://www.verizon.com/business/resources/T4d/reports/2024-dbir-data-breach-investigations-report.pdf>

A graph with numbers and text

Description automatically generated

Kuva 29 Tietomurroissa käytetyt menetelmät IBM:n tutkimuksessa

# yhteenveto

# Lähteet

1. https://www.dataversity.net/brief-history-cloud-computing/, viitattu 3.9.2024

1. <https://www.dataversity.net/how-the-cloud-has-evolved-over-the-past-10-years/>, viitattu 3.9.2024

1. <https://www.cloudzero.com/blog/cloud-computing-statistics/>, viitattu 3.9.2024

1. <https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-private-public-hybrid-clouds>, viitattu 3.9.2024

1. <https://www.marketsandmarkets.com/Market-Reports/cloud-storage-market-902.html>, viitattu 3.9.2024

1. <https://www.cloudflare.com/learning/ddos/glossary/hypertext-transfer-protocol-http/>, viitattu 4.9.2024

1. <https://httpwg.org/specs/rfc9110.html>, viitattu 4.9.2024

1. <https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview>, viitattu 4.9.2024

1. <https://developer.mozilla.org/en-US/docs/Web/HTTP/Status>, viitattu 4.9.2024

1. <https://www.javacodegeeks.com/2023/03/http-1-1-vs-http-2-vs-http-3-key-differences.html>, viitattu 4.9.2024

1. <https://ably.com/topic/http-2-vs-http-3>, viitattu 4.9.2024

1. <https://www.w3.org/Protocols/HTTP/1.0/spec>, viitattu 4.9.2024

1. <https://www.rfc-editor.org/rfc/rfc9114.txt>, viitattu 4.9.2024

1. <https://http.dev/3>, viitattu 4.9.2024

1. <https://www.cloudwards.net/what-is-webdav/>, viitattu 4.9.2024

1. <https://datatracker.ietf.org/doc/html/rfc4918>, viitattu 4.9.2024

1. <http://www.webdav.org/specs/rfc3744.html>, viitattu 4.9.2024

1. <https://developer.mozilla.org/en-US/docs/Web/API/WebSocket>, viitattu 5.9.2024

1. <https://datatracker.ietf.org/doc/html/rfc6455>, viitattu 5.9.2024

1. <https://http.dev/ws>, viitattu 5.9.2024

1. <https://tools.ietf.org/html/rfc959>, viitattu 5.9.2024

1. <https://www.rfc-editor.org/rfc/rfc4217.txt>, viitattu 5.9.2024

1. <https://www.ssh.com/academy/ssh/sftp-ssh-file-transfer-protocol>, viitattu 5.9.2024

1. <https://datatracker.ietf.org/doc/html/draft-ietf-secsh-filexfer-02>, viitattu 5.9.2024

1. <https://www.ibm.com/docs/en/ssw_ibm_i_71/sqlp/rbafy.pdf>, viitattu 5.9.2024

1. <https://www.postgresql.org/docs/current/routine-vacuuming.html>, viitattu 5.9.2024

1. <https://en.wikibooks.org/wiki/PostgreSQL/Architecture>, viitattu 5.9.2024
2. <https://www.mongodb.com/resources/basics/databases/document-databases>, viitattu 6.9.2024
3. <https://www.mongodb.com/resources/basics/databases/types>, viitattu 6.9.2024
4. <https://www.mongodb.com/resources/basics/databases/nosql-explained>, viitattu 6.9.2024

|  |
| --- |
|  |