DataLake-hanke - AWS

26.9.2018, tietohallinto, Matti Valli, Pasi Porkka

DataLake – Organisaation näkökulma

Yhteen järjestelmään kerätty organisaation liiketoiminnallinen tieto parantaa organisaation mahdollisuuksia saada ajantasaista ja kattavaa tietoa sen omasta liiketoiminnasta:

- Kaikilla toimijoilla on samat tunnusluvut, jotka perustuvat samaan tietoon ja samoihin analyyseihin
- Tieto on jatkuvasti ajan tasalla
- Useista eri operatiivisista järjestelmistä saatua tietoa voidaan yhdistellä keskenään ja muun esim. avoimen datan ja paikkatiedon kanssa
- BI-alustalla vaativat analyysit ovat tehokkaita ja helpohkosti toteutettavissa
- Kullekin toimialalle voidaan räätälöidä omia raportteja
- Henkilöstö voi itse analysoida koostettuja tietoja sallituilla ohjelmistoilla
- BI-alusta mahdollistaa simulaatiot ja aikasarjaennusteet
- Asiakkaille (kaupungin asukkaille) voidaan tehdä omia näkymiä

DataLake – Lähtökohta ICT-yksikkö

- Vuonna 2017 n. 210 eri tietojärjestelmää, joissa osassa tietokanta, osassa ei
- Kaikkiin kantoihin ei kuitenkaan saatavilla rajapintaa; (uusissa sopimuksissa pykälä, joka takaa pääsyn kantaan eli omaan dataan!)
- Omien järjestelmien lisäksi avointa dataa
  - Väestörekisteri, verottaja, PRH, kaupparekisteri, jne.
- Data saatavilla useissa erilaisissa muodoissa (csv, xml, json, txt, doc, pdf, jne.)
- Isoimmissa operatiivisissa tietokannoissa satoja tauluja
- Tietomallit monimutkaisia; esim. HR-puolella tietokuutio, jossa 36 dimensiota
- Operatiivisissa järjestelmissä kymmeniä omia raportointiohjelmia

DataLake – Tavoitteet ICT-yksikkö

- Koska paljon erilaisia datalähteitä:
  - Erilaisille datatyypeille omat automatisoidut datan siirtoputket.
  - Toteutus ja suorituksen automatisointi yhdenmukaisesti.
  - Integraatioiden on reagoitava, jos toipumisen automatisointi ei ole järkevää. Data ei saa korruptoitua automaation seurauksena!
  - Oleellisen ja tärkeän datan määrittely priorisoiduista järjestelmistä.
- Koska data monimutkaista:
  - Dokumentaatio – data, prosessit, roolit, päätökset, integraatiot ja erityisesti riippuvuudet eri tietolähteiden välillä on kuvattava jatkuvasti; **kokonaiskuva**
- Koska (toivottavasti) käytetään vain muutamaa raportointiohjelmaa:
  - raporttien yhtenäiset pohjat, yhtenäinen laskenta, sama värimaailma, yhteiset tulkinnat
  - raportointiohjelmien syvällinen osaaminen ja koulutus toimialoille

DataLake - Poc

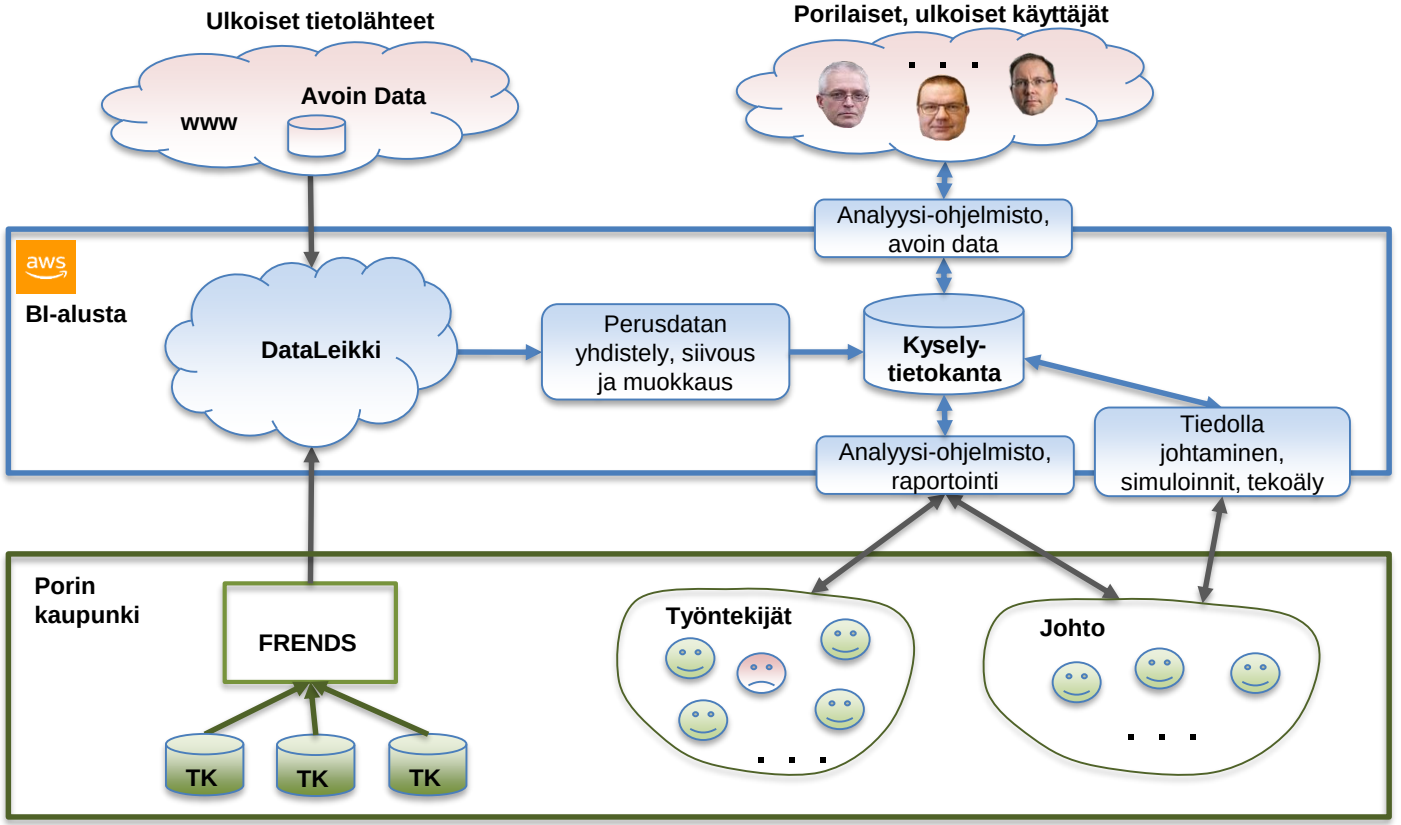
- 2016 alkaen testattiin samalla aineistolla 3:n eri teknologiaratkaisun ETL-prosessia
- Aineistoina ostolaskutiedot, varastotapahtumat, maaomaisuus ja kaivuuluvat

Omia kokemuksia ja mielikuvia toimittajista ja tuotteista:

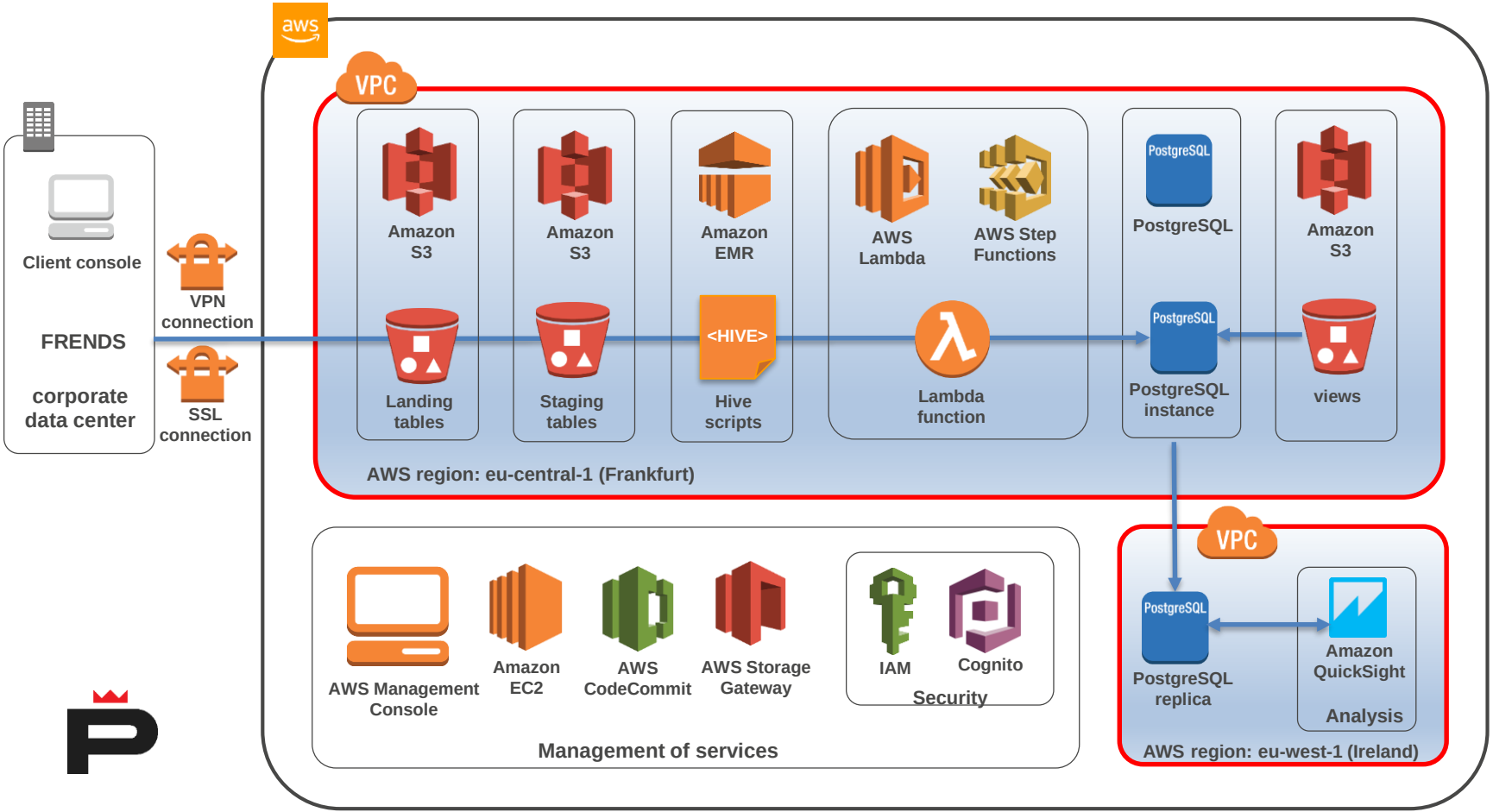
- **Microsoft** – sekava arkkitehtuuri, liikaa liikkuvia osia, hintapolitiikka, ei kustannusarviota
- **IBM** – pystytysvaikeudet, aikataulu petti, luottamuspula
- **Oracle** – selkeä arkkitehtuuri, kallis, toteutui halutusti, hintapolitiikka?
- Suunnitelman mukaisesti toteutui Oracle, mutta hyvin kallis. Microsoftkin ehkä (!?).
- IBM-alustan kanssa suuria teknisiä pystytysvaikeuksia eikä sen kanssa päästy eteenpäin.
- Kysyttiin myös **Fujitsu**n tarjonta DataLaken pohjaksi → ei oikein sopinut ajatuksiin

→ kypsyi ajatus kokeilla arkkitehtuuria, joka mahdollistaa
open source- pohjaisten BI- ja analytiikka-työkalujen käytön
→ **AWS!**

DataLake – Yleisarkkitehtuuri



DataLake – AWS - arkkitehtuuri



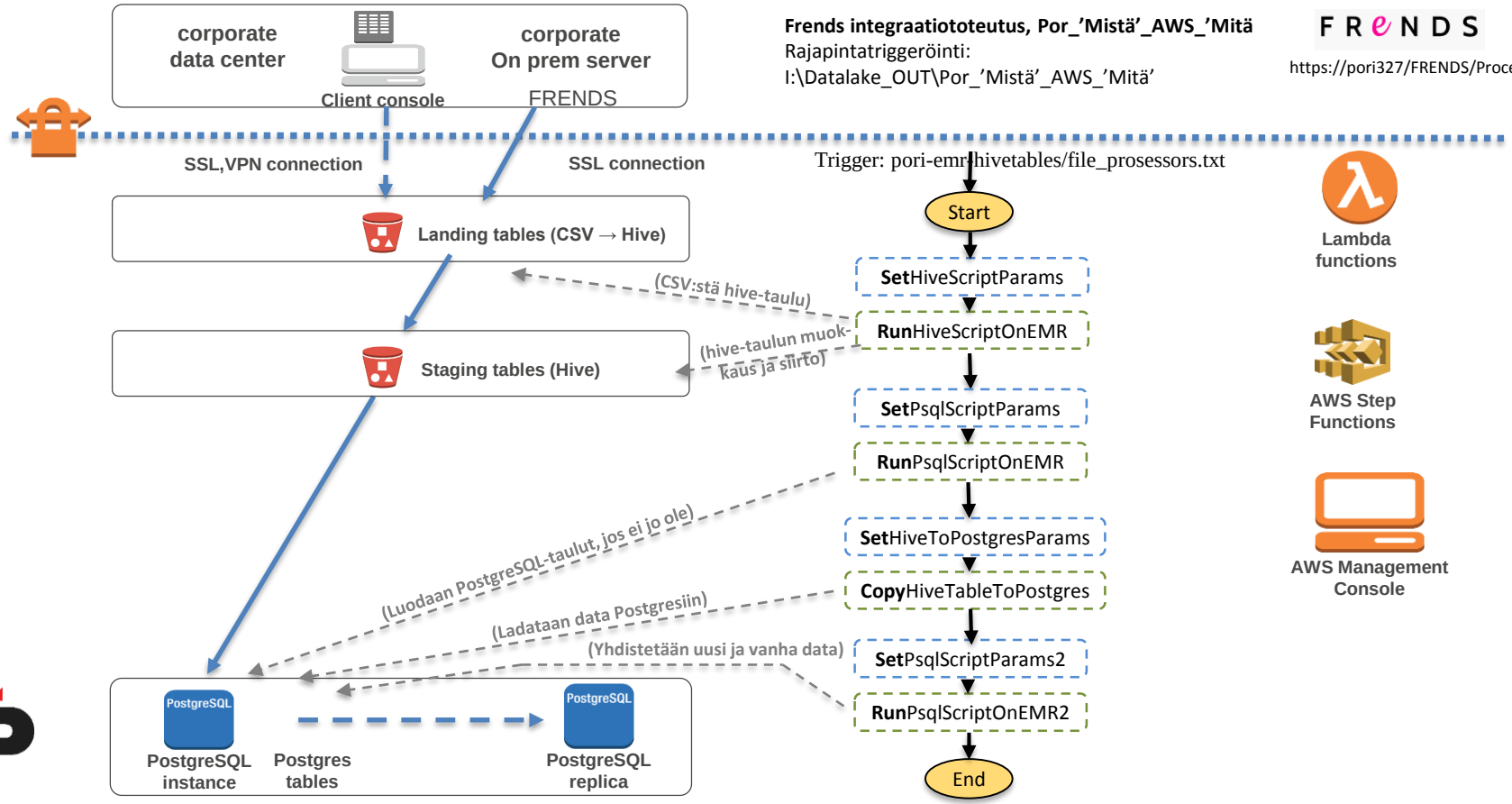
DataLake – AWS - komponentit

- Palveluja otetaan käyttöön **AWS Management Console**n avulla, joka on ainoa pakollinen AWS:n komponentti. Alla on esitelty niitä palveluita, joita tässä pilotissa käytetään.
- Järjestelmän hallinnointipalvelut
  - **IAM** – Identity and Access Management, käyttöoikeuksien hallinnointi
  - **Cognito**, käyttäjäidentiteetin hallintapalvelu mobiilisovelluksille
  - **EC2 –** Elastic Compute Cloud, tietoturvallinen ja skaalautuva pilven ohjelmointialusta
  - **VPC** – Virtual Private Cloud, pilviresurssien eristyspalvelu
-
- Tallennuspalvelut, niiden hallinta ja tietokannat
  - **S3** – Simple Storage Service, skaalautuva tallennustilapalvelu
  - **Storage gateway**, oman IT-ympäristön integrointi AWS:ään
  - **RDS** – Relational Database Service, relaatiotietokannan käyttöönotto
- Sovellusten kehitys, koodin hallinta ja automatisointi.
  - **Lambda**, tapahtumapohjainen informaationkäsittelypalvelu
  - **CodeCommit**, koodin versionhallinta
  - **AWS Step Functions**, sovellusten ja mikropalvelujen hallinta visuaalisin työkaluin
  - **Amazon EMR**, Hadoop viitekehys hive-funktioiden käytölle ja HUE-käyttöliittymä
- Analyysi ja raportointi
  - **Amazon QuickSight**, integroitu analyysien ja visualisoinnin työkalu

Siirron vaiheet

Siirron prosessi

Hallinta



DataLake - roadmap

- [x] AWS:n käyttöönotto, perustoiminnallisuus, tallennusalueet, postgreSQL
  Erimuotoisen datan putkitus datalakeen ja postgresiin
  - [x] CSV
  - [x] XML
  - [ ] …
- [ ] Avoimen datan siirrot, tallennusrakenteet ja bucketit. Menetelmät!
- [ ] Ei-rakenteisen datan tallennusrakenteet S3:ssa
- [ ] Analytiikka- ja aikasarja-analyysien ”tuotteistus”
- [ ] Hakukoneen (Cloud Search) käyttöönotto
- [ ] Oma sovelluskehitys AWS:ssä
- [ ] Haaveena OmaData-palvelu kansalaiselle
- [ ] …