Big data

Lasse Seppänen

Erilaisia datoja

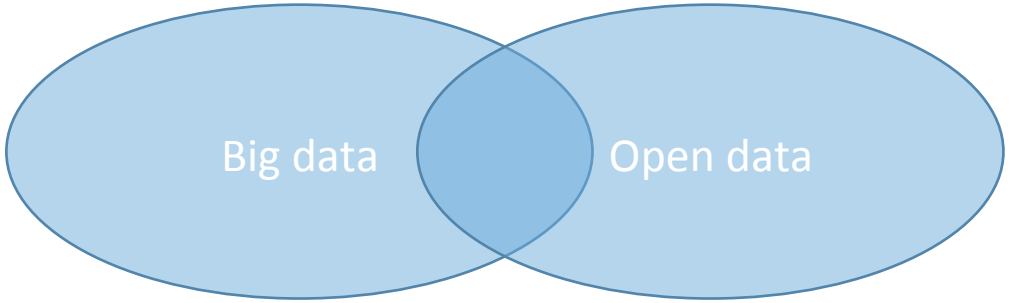
- Kaikki data on pohjimmiltaan bittejä.
  - Data on digitaalista raaka-ainetta, kuten tilastoja, taloustietoja, karttoja, kuvia, videotallenteita ja 3D-malleja.
  - Erilaiset datat sopivat silmälle ja tietoteknisesti prosessoitaviksi
    - kuvat vs. relaatiokanta
- Yrityksen toiminnanohjausjärjestelmän ja muiden järjestelmien datat
  - Hyvin organistoitua dataa
  - SQL selecteillä haut dimensioiden mukaan (Business Intelligence)
- Teollisuuden koneiden anturien tuottama data

Big data

- Erittäin suurten, järjestelemättömien, jatkuvasti lisääntyvien tietomassojen keräämistä, säilyttämistä, jakamista, etsimistä, analysointia sekä esittämistä tilastotiedettä ja tietotekniikkaa hyödyntäen
- Ei täysin vakiintumatonta määritelmää, mutta
  - se ei ole käsiteltävissä yleisesti käytössä olevilla laitteistoilla tai ohjelmistoilla siedettävissä olevassa ajassa käyttäjän kannalta
  - mahdollisesti käytössä monessa paikassa yhtä aikaa
  - data tulee eri lähteistä, eri muodoissa ja se kasaantuu ja/tai muuttuu nopeasti
  - usein jonkin laitteen automaattisesti tuottamaa
  - kerätty mahdollisesti ilman suunnitelmaa siitä, mihin sitä tarkkaan ottaen tullaan käyttämään
  - datalla on usein vain löyhästi määritelty rakenne, tai ei rakennetta lainkaan, jolloin sitä ei voida sellaisenaan analysoida

wikipedia

Avoimet datat



- **Julkisuus**: Datan on sisällettävä julkista tietoa.
  - Yksityisyydensuoja ja yleinen turvallisuus
  - Ei henkilötietoja tai liikesalaisuuksia.
- **Tekninen saatavuus**:
  - muodossa, jotta on käsiteltävissä tietokoneella.
    - Ihmiselle helppoja PDF-dokumentit (Readerilla) tai HTML-sivut (selaimen jälkeen), mutta niitä on vaikea tulkita ohjelmallisesti.
    - Koneellisesti sopivat esim. CSV-, XLS- tai XML-muodot sekä erilaiset rajapinnat suoraan datalähteeseen.
- **Maksuttomuus**
- **Uudelleenkäytön sallivat käyttöehdot**:
  - datan yhteydestä löytyvät käyttöehdot
- Ilmatieteen laitoksen avoimet datat

Big datan määrä ja operaatiot

- Milloin datan määrä on niin iso, että sitä voidaan ajatella big dataksi?
- Analyysimenetelmiä voidaan käyttää myös pienempien tietomäärien analysoimiseksi.

Dataluokituksia

- Rakenteellisuus
  - Selkeä ja ennalta tarkkaan määritelty rakenne. Perinteisesti data on tällaista, missä on ennalta määritetty mitä tietoja kerätään ja miten ne merkitään ja tämä sama säännöstö pätee koko dataan.
  - Löyhästi määritelty rakenne. Esimerkiksi internetsivuston keräämät lokitiedot ovat tällaista. Analysointia varten dataa joudutaan luultavasti merkittävästi muokkaamaan ja sieltä poimimaan hyödylliset osat.
  - Ei lainkaan rakennetta. Esimerkiksi asiakaspalautteet tai sosiaalisesta mediasta poimitut ihmisten lähettämät julkiset päivitykset. Tällaisessa tilanteessa ei voida lainkaan tehdä oletuksia siitä, mitä data tulee sisältämään tai millaisessa muodossa asioita tullaan ilmaisemaan.
- Turha / hyödyllinen data

Englanniksi v-sanoja

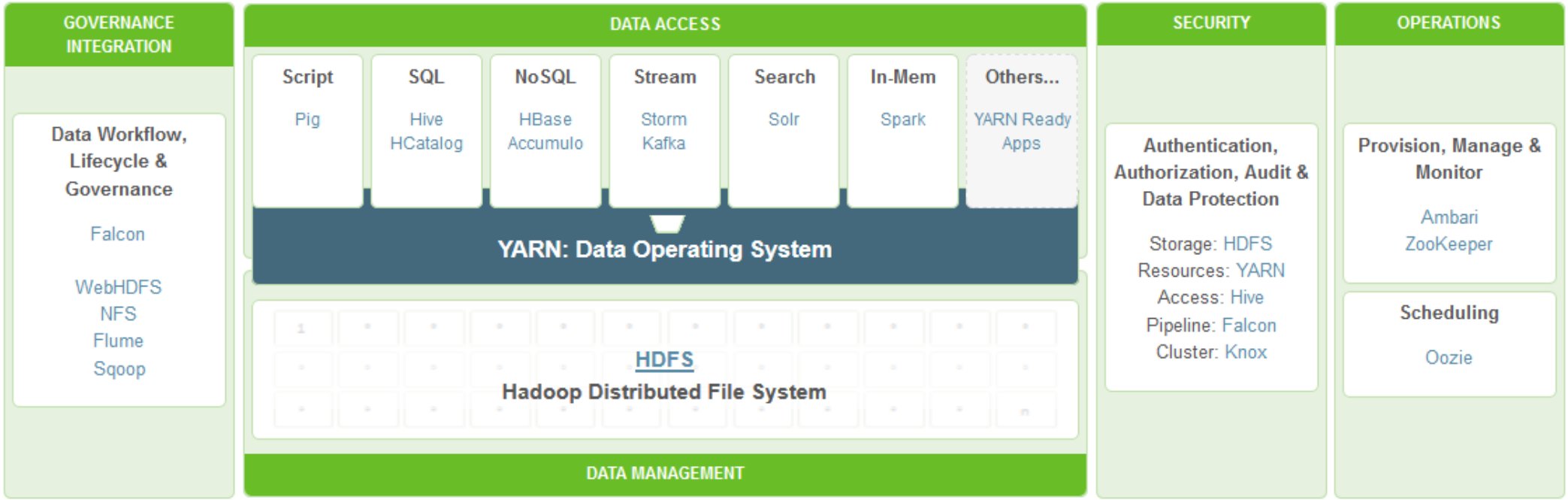
- Volume
- Velocity
- Variety

- Variability
- Complexity

Big data technologies: Hadoop

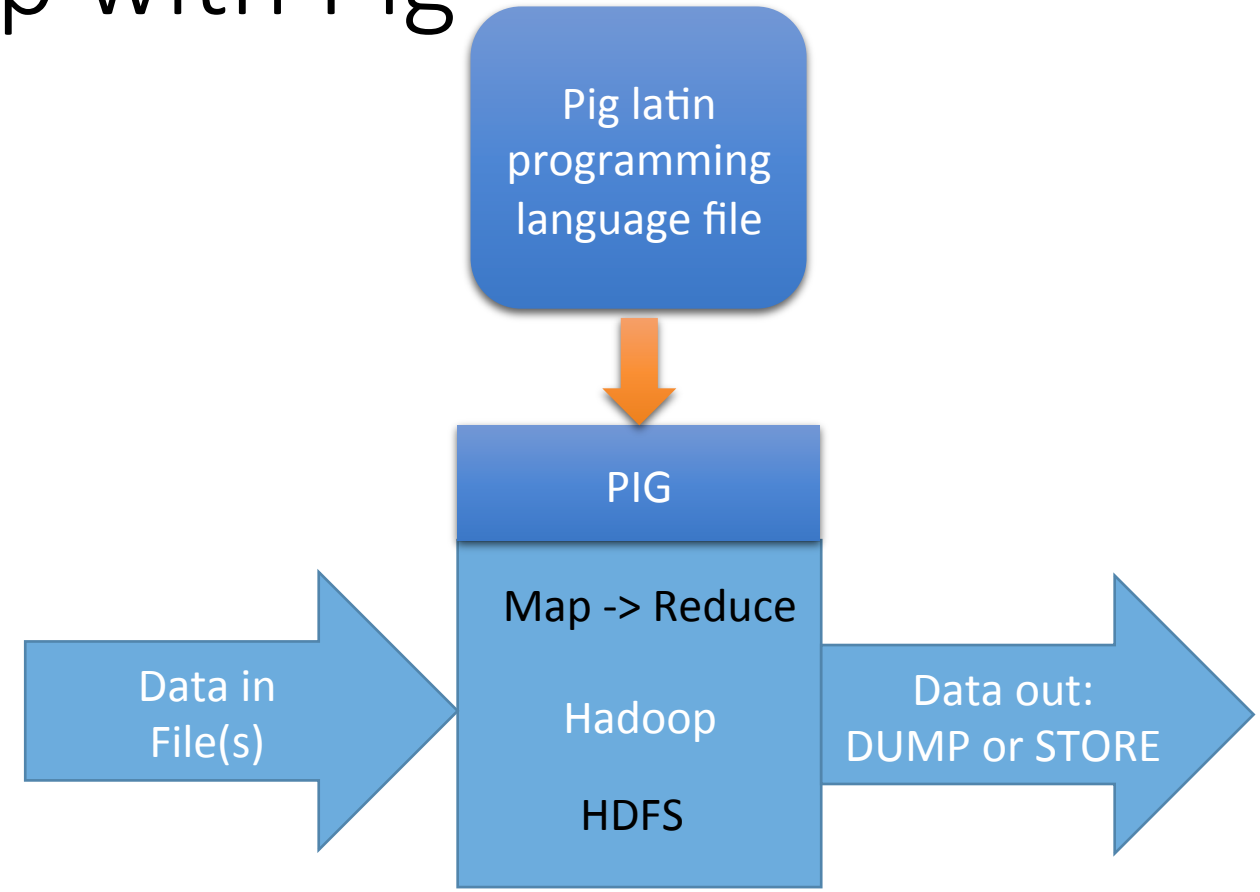
- **Hadoop** the most popular implementation of MapReduce, an open source platform for handling Big Data. Works with multiple data sources.
- **MapReduce** is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. MapReduce implementation consists of two tasks:
  - The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples;
  - The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).
- **Hive** is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. Allows anyone to make queries to data stored in a Hadoop cluster just as if to a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.
  - Developed by Facebook, now open source.
- **PIG** is a bridge that tries to bring Hadoop closer to the realities of developers and business users. PIG consists of a Pig latin language that allows for query execution over data stored on a Hadoop cluster.
  - Developed by Yahoo!, now open source.
- A lot of Hadoop modules.

The whole Hadoop picture



Source: Hortonworks

Hadoop with Pig



Pig latin syntax

- <variable name> = <some action>

- <variable name> can be used in next <actions>.

Used files

- STOCK_A = LOAD 'NYSE_daily_prices_A.csv' using PigStorage(',') AS (exchange:chararray, symbol:chararray, date:chararray, open:float, high:float, low:float, close:float, volume:int, adj_close:float);
- STOCK_B = …

- <Actions>

- DUMP <action> to the screen
- STORE <action> INTO a file

Example

- batting = load ’Batting.csv`using PigStorage(’,’);
- runs = FOREACH batting GENERATE $0 as playerID, $1 as year, $8 as runs;
- grp_data = GROUP runs BY (year);
- max_runs = FOREACH grp_data GENERATE group as grp, MAX(runs.runs) as max_runs;
- join_max_run = JOIN max_runs by ($0, max_runs), runs by (year,runs);
- join_data = FOREACH join_max_run GENERATE $0 as year, $2 as playerID, $1 as result;
- DUMP join_data;