Data Engineering
Me

- Pietari Kettunen
- MSc in Software Engineering
- Working in software since 2003
- Data “something” since about 2014/15
- @Pietrotull
Sqooba

- Data Company
- Founded 2016
- Based in Bern
- 20+ employees
Customers

- AXA
- PostFinance
- T.. (TUI AG)
- Sunrise
- die Mobiliar
- Schindler
- ACMEIRA
- Givaudan
Agenda

1. What is Data Engineering
2. Working community around data
3. what is data
4. Data Pipelines
5. Tools
6. Functional Programming
7. Challenges
Software Engineering

“Software Engineers design, construct, test, and maintain computer applications software or systems to meet the needs of stakeholders.”

Data Engineering

“Data Engineers build pipelines that transform and move data into formats and places so that data scientists can work on it”

Plumbing
Working in Data

Data Scientist

Data Engineer

DevOps
“Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.”
Data Scientist

- Uses the data, works on the data
- Analyse, identify patterns
- Visualizations
- Build models, make predictions

*Example: “predict churn rate”*
DevOps

- Provides infrastructure to run applications
  - required software tools / platforms
  - security
- Monitor cluster
- Maintain infrastructure
- Supports data engineers
Data Engineering

- Provides Data Scientists with data
- Subset of software engineering
- Connecting data scientists with data
- Dealing with large amount of data (Big Data)
- Ingesting data and building data pipelines
- work closely with DevOps especially early stages
Aim at clean plumbing
What is data?

- Database
  - transaction logs
- Log files
- Event streams
- Data comes in many forms and formats
Database is “snapshot”

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Balance</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>John Smith</td>
<td>+ 5.15</td>
<td>Balance</td>
</tr>
<tr>
<td>101</td>
<td>Bob Smith</td>
<td>+ 5.15</td>
<td>Balance</td>
</tr>
<tr>
<td>Id</td>
<td>Name</td>
<td>Balance</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>---------------</td>
<td>---------</td>
<td>--------------</td>
</tr>
<tr>
<td>100</td>
<td>John Smith</td>
<td>+ 5.15</td>
<td>Balance</td>
</tr>
</tbody>
</table>
What is data?

- database is a "snapshot"
- logs tend to be more interesting
  - state can be reconstructed
  - more granularity
- Time series
Data Pipeline

- Cleaning
- Transforming
- Formatting
- Moving
- Storing
Batch vs Stream

- Stream is a continuous flow
  - a lot of hype
- Batch data comes in blobs
  - still the more reasonable option
- extract, transform, load
  - extract from database / data source
  - transform data as necessary
  - load into another database
Data Warehouse vs Data Lake

- Proprietary enterprise solution for data analysis.
  - tends to get very expensive over time

- Data Lake usually build on open source tools
  - flexible
  - requires more knowhow
  - lacking in support
Tools

- **Kafka / streaming platform**
- **Apache NiFi**
  - Graphical data engineering tool
- **Apache Spark**
  - Distributed general purpose cluster computing framework.
- **Apache Flink/Storm/and many**
  - Stream processing framework
- **Hadoop (YARN, HDFS)**
Kafka

- Publish / subscribe message broker
- Highly scalable (10 to 1000 nodes)
- Created by LinkedIn
Spark

- Does everything "data"
  - sql and hive integration
  - machine learning
  - streaming
  - REPL

- "default" data processing engine

- REPL is great for data exploration / development
Spark

- RDD - Resilient Distributed Dataset
  - immutable
  - lazy
- everything is in memory
- Since version 2.x
  - Dataset
  - DataFrames
Spark WordCount example in Scala

```scala
val textFile = sc.textFile("hdfs://...")
val counts = textFile.flatMap(line => line.split(" "))
  .map(word => (word, 1))
  .reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://...")
```
Spark WordCount example in Java

```java
JavaRDD<String> textFile = sc.textFile("hdfs://...");
JavaPairRDD<String, Integer> counts = textFile
    .flatMap(s -> Arrays.asList(s.split(" ")).iterator())
    .mapToPair(word -> new Tuple2<>(word, 1))
    .reduceByKey((a, b) -> a + b);
counts.saveAsTextFile("hdfs://...");
```
Hadoop

- **HDFS (Hadoop Distributed File System)**
  - suitable for storing large amount of data

- **YARN (Yet Another Resource Negotiator)**
  - resource management for hadoop cluster

- **Hive**
  - Database that works on top of HDFS
  - JDBC driver
Graphical UI Tools Falling out of favour

- Apache NiFi
- Apache Kylo

"Best abstraction for software is code"
Functional Programming

- Avoid state
- Pure function
  - same function should always give the same output for the same input
- Reproducibility
- Scales
- Immutable
- Great match for Data Engineering
Data Engineering languages

- Scala
- Java
- (Python)
- (Clojure)
Example Architecture

Data Source

Kafka Connect

Real Time application

Kafka Cluster

Data Source

Data Source

Spark Application

Spark Application

Hadoop / HDFS / Hive - Cluster
Challenges

- Formats are many and they change
- 3rd party components, can be exotic
- Data type conversion
- Hard to make progress visible
- Scaling
Collaboration

- A data engineer alone can not do much
  - Data scientists are our customers
    - The data pipelines we build are actually designed to support data scientist work
  - DevOps are our partners
    - In all development steps, we need a DevOps team to enable the pipeline components (service installation, DNS configuration, FW, monitoring)
Questions?