Data Engineering Process Overview
Layers and Technologies

Netflix, Amazon, Zalando and Co are dealing with huge amounts (Petabytes) of data - every day. In order to deal with this new challenge technologies are being invented. Here is an overview of a set of frameworks and products that are currently used in data engineering processes. Disclaimer: This collection does by no means provide a whole picture of the available technologies. It rather presents a choice of common tools and categorises them into the data engineering process phases.

Analytics & Business Intelligence Layer & Machine Learning

This layer comprises technologies, that are able to turn complex data in more **accessible and understandable data** which can be studied, analysed and questioned by anyone. For instance, in this layer business information is analysed in order to make business decisions.

Tableau

- Powerful **data visualization tool**
- Used in Business Intelligence Industry
- Helps in simplifying raw data into an easily understandable format (charts …)
- Features include:
  - Data Blending
  - Real Time Analysis
  - Collaboration of Data

Vizdom & IDEA

- Vizdom: **interactive data visualisation tool**
  - Allows pen and touch interactions to create visualizations
- IDEA: interactive data exploration accelerator
  - Operates as a middleware between the data source and the Vizdom application using an approximate query engine
  - Splits SQL-queries into smaller queries, approximates result, and uses a result cache based on Bayes Theorem to cache intermediate results of all visualization as random variables

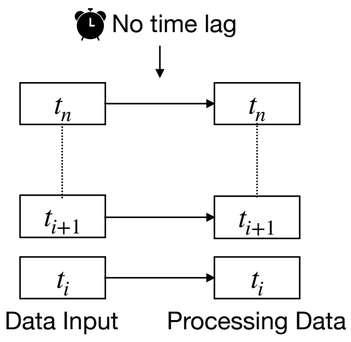
TPOT

- **Automated machine learning** (AutoML) software
- Explores thousands of possible pipelines to find the best for a given data set
- Uses genetic programming
- Once it finished searching, it provides the user with the Python code for the best pipeline

Stream Processing Technologies

Data Processing & Query Layer

These technologies can take data, join different sets, reduce it to key-value pairs, and then run calculations on adjacent pairs to produce some final calculated value.
Data items can also be plugged into machine learning algorithms to make some projection (predictive models) or discover patterns (classification models).
Here I am focusing on technologies that are able to perform **stream processing.** In stream processing, we process data as soon as it arrives in the storage layer - which is often very close to the time it was generated.
Example use case: detect anomalies that signal fraud in real time, then stop fraudulent transactions before they are completed.



Data and result storage

Apache Flink

- Open-source cluster for streaming and processing data
- **Stream processing**
- Fault-tolerant and built to scale to immense amounts of data
- Cross platform and supports most of the application integrations
- As far as streaming capability is concerned Flink is faster than Spark, since Spark handles stream in form of micro-batches and has no native support for streaming

vs.

Apache Spark

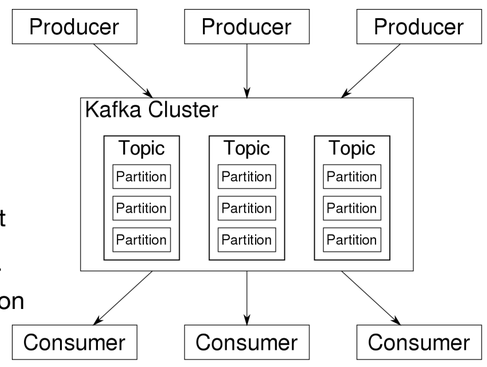
- Open-source cluster for big data processing
- **Stream and Batch processing**
- Fault-tolerant and built to scale to immense amounts of data
- Operated using third party cluster managers
- Spark has very strong community support and has a good number of contributors.

Data Integration & Ingestion Layer

Information coming in in real-time can be processed in a stream and stored afterwards. This layer is the first step for the data coming from variable sources to start its journey. Data here is **prioritised and categorised** which makes data flow smoothly in further layers.

Apache Kafka

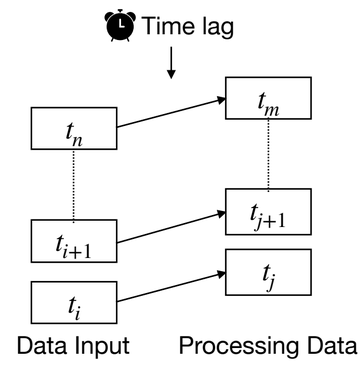
- Open source **distributed streaming platform**
- 3 Key capabilities:
  - publish and subscribe to streams of records, similar to a message queue or enterprise messaging system
  - store streams of records in a fault-tolerant durable way
  - process streams of records as they occur
- Mature and powerful solution used in production at huge scale



Batch Processing Technologies

Data Warehousing

Typical ETL based data warehousing allows to **extract, transform and load** data:
- Extract: originating from various different sources the data has to be gathered in order to be persisted and analysed
- Transform: traditional data warehouses use **batch processing**. This means, that newly arriving data elements are collected into a group. The whole group is then processed at a future time (as a batch, hence the term “batch processing”).
- Load: the access layer helps to retrieve data



Hadoop HDFS

- Classic **big data file system**
- Popular due to its robustness and limitless scale
- Files are stored on multiple machines redundantly to resume the system from possible data losses in case of failure
- Makes applications available for parallel processing

Hive

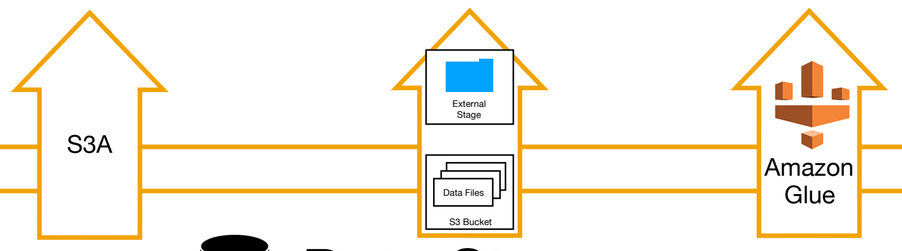
- **Data Warehouse** for Hadoop
- Makes Hadoop cluster feel like a relational database
- Allows you to write SQL (actually HiveQL) queries against data stored in HDFS

Snowflake

- **Data warehouse built for the cloud**
- Data stored on Cloud Storage
  - Partitioned into micro partitions
  - File metadata helps with pruning, allows time travel, cloning
- Automatic clustering
  - Technique to optimise data layout in the background
  - Maintain an approximate sort order of the data
  - Incrementally re-cluster batches of files, selected to minimize the depth

Amazon Redshift

- **Petabyte-scale data warehouse service** in the cloud
- Known its processing speed
- Column-oriented database designed to connect to SQL-based clients and BI tools
- Use cases:
  - Traditional Data Warehousing
  - Store & Process Data with log analysis
  - Analyse Data for business applications
  - Time-sensitive data reporting for mission-critical workloads



Data Storage

Amazon S3 (Amazon Simple Storage Service)

- **Object storage service** with a simple web service interface to store and retrieve any amount of data from anywhere in the internet
- 99,9% durability and availability
- Use Case:
  - Storage for cloud-native applications
  - Backup and Disaster Recovery
  - Bulk repository or “data lake” for analytics
  - Data Archive

Continuous Real-Time Data

stock ticker prices, weather readings coming from sensors, temperature and weather gauges on industrial machines and even tweets (Twitter has API to query tweets in real time)

Sources:

Presentations:
- Towards Interactive Data Analytics
- File Metadata Management in Snowflake
- Apache Flink - An Introduktion and Outlook into the Future

Websites:
- https://www.tableau.com
- https://epistasislab.github.io/tpot/
- https://spark.apache.org
- https://hadoop.apache.org
- https://aws.amazon.com/de/redshift/
- https://aws.amazon.com/de/s3/

Angelika Wieck
Bachelor IT-Systems Enginerring
5th semester

Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam
E-Mail: angelika.wieck@student.hpi.de

Hasso Plattner Institut
Digital Engineering · Universität Potsdam