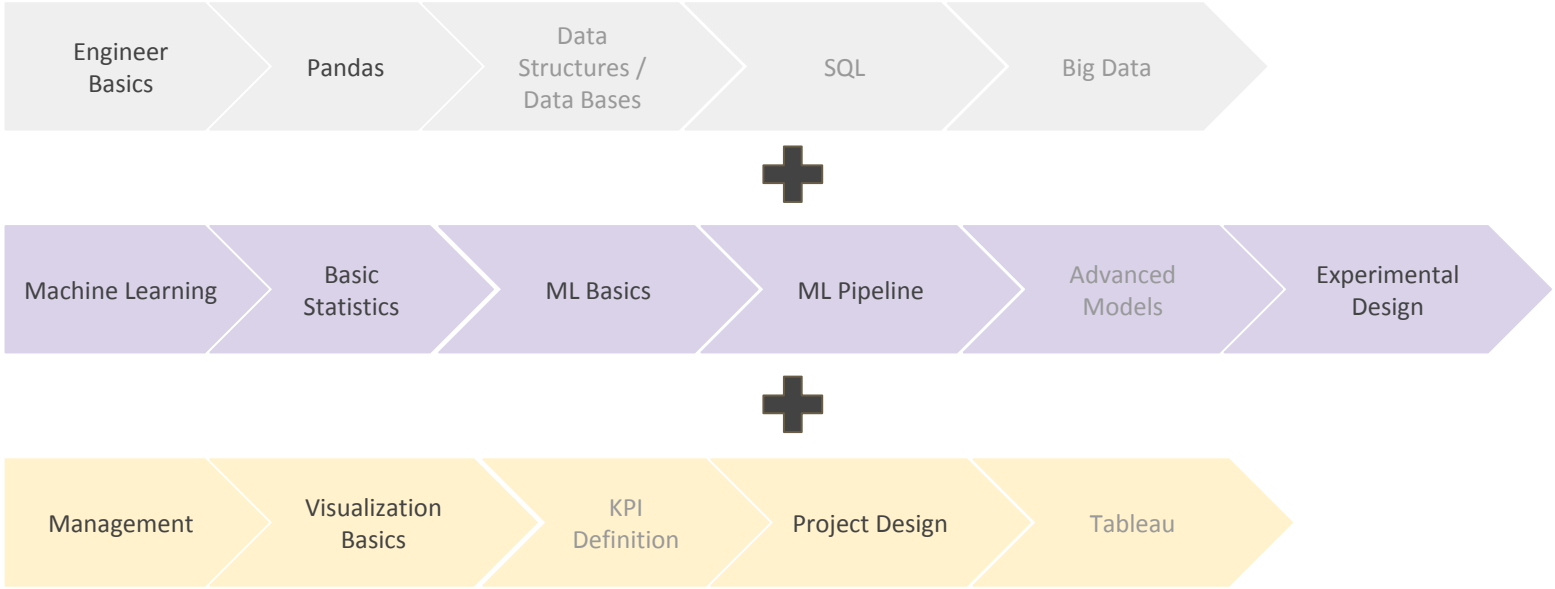
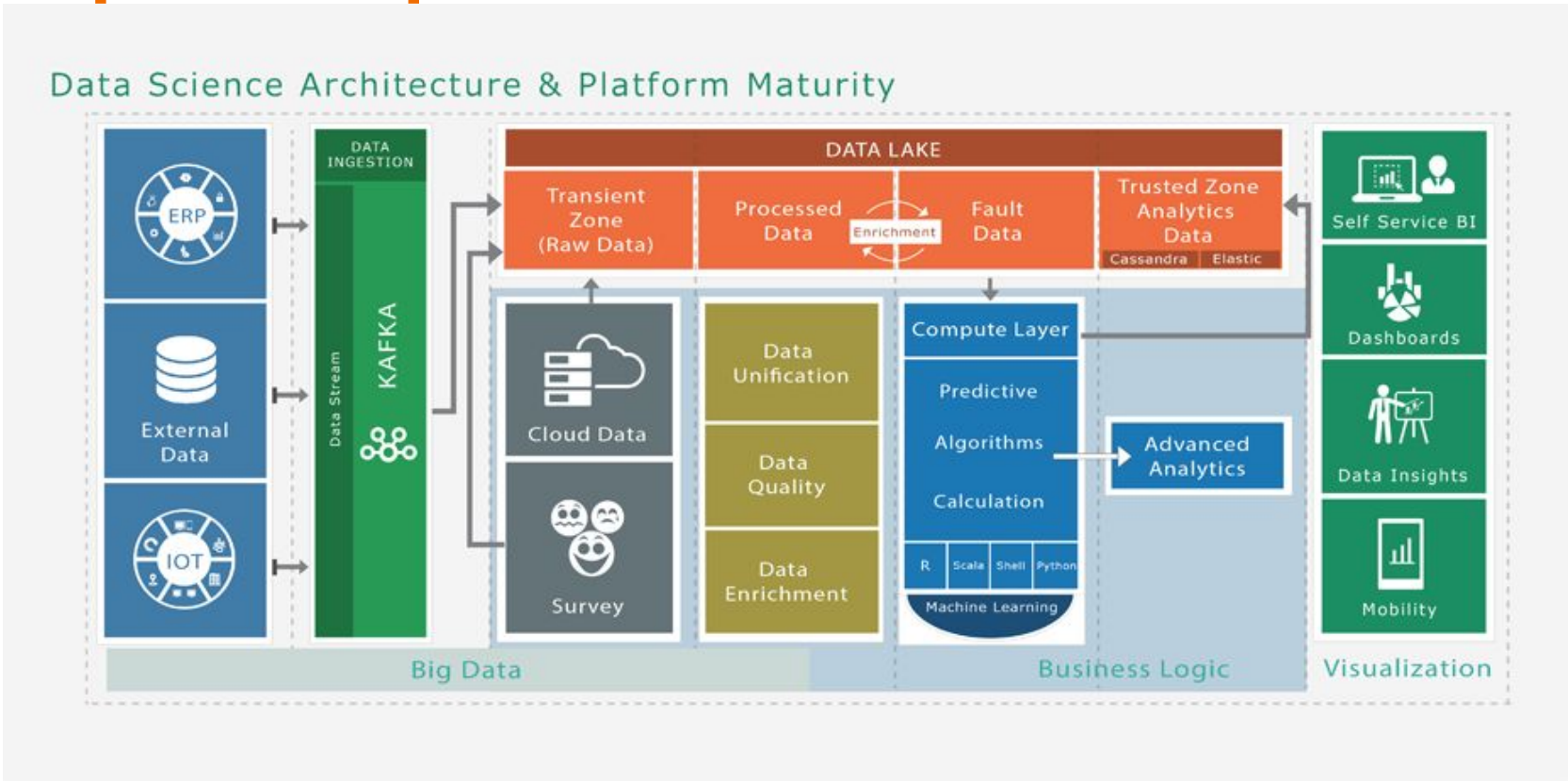
Technology Landscape

Data Engineering

Course Overview



Recap: Data Pipeline



Technologies: Data Ingestion

Big data ingestion is about moving data - especially unstructured data - from where it is originated, into a system where it can be stored and analyzed.

It may be continuous or asynchronous, real-time or batched or both (lambda architecture)

Streaming or Batch?

Technologies:

- Apache Kafka
- Amazon Kinesis
- Apache Flume

Data Ingestion: Apache Kafka

Name: Apache Kafka (kafka.apache.org)
Vendor: Apache Software Foundation (Open Source)
Time in the market: ~7 years

Popular Use Cases:

- Building real-time streaming data pipelines that reliably get data between systems or applications
- Building real-time streaming applications that transform or react to the streams of data

Brief description:

- 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

Alternatives: none for the entire stack. Some alternative queuing systems: ZeroMQ, ActiveMQ and RabbitMQ

Data Ingestion: Amazon Kinesis

Name: Amazon Kinesis (https://aws.amazon.com/kinesis/)
Vendor: Amazon (Proprietary)
Popular Use Cases:

- Netflix: almost realtime application monitoring

Brief description:
Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.

Alternatives: Apache Kafka

Data Ingestion: Apache Flume

Name: Apache Flume (flume.apache.org)
Vendor: Apache Software Foundation (Open Source)
Popular Clients: Bloomberg, Reuters, CMSWire: ingestion of news content

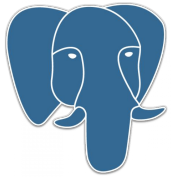
Brief description:
Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

Alternatives: Elasticsearch/Logstash/Kibana (ELK), Kafka

Technologies: Data Storage

- PostgreSQL
- Apache Cassandra
- Redis
- Apache HBase
- InfluxDB
- Neo4j

Data Storage: PostgreSQL



Name: PostgreSQL (www.postgresql.org)
Vendor: PostgreSQL (Open Source)
Time in the market: +25 years
Popular Clients: TripAdvisor.com, Instagram, Reddit, Skype, OpenStreetMaps

Brief description:

- PostgreSQL is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards compliance.
- Its primary functions are to store data securely and return that data in response to requests from other software applications

Alternatives: MySQL, SQL Server, Oracle

Data Storage: Apache Cassandra

Name: Apache Cassandra (cassandra.apache.org)
Vendor: Apache Foundation (Open Source)
Time in the market: ~8 years
Popular Use Cases:

- Apple uses 100,000 Cassandra nodes
- CERN used Cassandra-based prototype for its ATLAS experiment
- Netflix uses Cassandra as their back-end database for their streaming services

Brief description:

- A distributed NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.

Alternatives: Google BigTable, Amazon DynamoDB, CouchDB, MongoDB

Data Storage: Redis

Name: Redis (www.redis.io)
Vendor: Redis Labs (Open Source)
Time in the market: ~8 years
Popular Clients: Twitter, GitHub, Pinterest, Snapchat, Craigslist, StackOverflow, Flickr

Which problem can be solved:

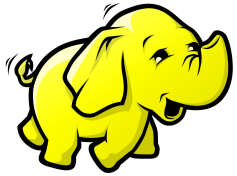
- Index a large dataset in realtime

Brief description:

- In-memory data structure store, used as a database, cache and message broker
- It supports data structures such as strings, hashes, lists, sets, sorted sets with range queries, bitmaps, hyperloglogs and geospatial indexes with radius queries

Alternatives: Aerospike, MemcacheDB, Amazon ElasticCache, Azure Cache

Data Storage: Apache HBase



Name: Apache HBase (hbase.apache.org)
Vendor: Apache Software Foundation (Open Source)
Popular Use Cases:

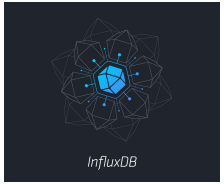
- Amadeus IT Group, as its main long-term storage DB
- Airbnb uses HBase as part of its AirStream realtime stream computation framework
- Netflix, Pinterest, Yahoo!

Brief description:

- HBase is an open-source, non-relational, distributed database modeled after Google's Bigtable and is written in Java

Alternatives: Google BigTable, Cassandra

Data Storage: InfluxDB



Name: InfluxDB (www.influxdata.com)
Vendor: InfluxData (Open Source)
Time in the market: ~4 years
Popular Use Cases:

- BBOXX uses InfluxData as its IoT Data Platform, to continuously monitor their geographically dispersed 85,000 solar units
- IBM uses InfluxData to provide monitoring, visibility, and control of its Trusteer product line

Which problem can be solved: Retrieval of time series data
Brief description:

- Time series database
- Optimized for fast, high-availability storage and retrieval of time series data in fields such as operations monitoring, application metrics, Internet of Things sensor data, and real-time analytics

Alternatives: Riak-TS, Graphite

Data Storage: Neo4j

Name: Neo4j(www.neo4j.com)
Vendor: Neo4j, Inc (Open Source)
Popular Use Cases:

- Walmart: quickly query customers' past purchases, and instantly capture any new interests (essential for making real-time recommendations)
- E-Bay

Which problem can be solved: Store and represent data relations in a convenient way
Brief description:

- A high performance graph database management system
- Includes all the features expected of a mature and robust database

Alternatives: AllegroGraph, Apache Giraph

Technologies: Data Processing

- Apache Spark
- ElasticSearch

Machine Learning tools:

- TensorFlow

Data Processing: Apache Spark

Name: Apache Spark (spark.apache.org)
Vendor: Apache Software Foundation (Open Source)
Time in the market: ~3 years
Popular Clients: Oracle, Hortonworks, Cisco, Visa, Microsoft, Databricks and Amazon
Which problem can be solved: Processing a large amount of data in batch and realtime
Brief description:

- An open-source cluster-computing framework
- Has as its architectural foundation the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way
- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk
- Combine SQL, streaming, and complex analytics
- Spark runs on Hadoop, Mesos, Kubernetes, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3

Alternatives: Hadoop, Apache Storm, Apache Flink

Data Processing: ElasticSearch

Name: ElasticSearch (elastic.co/products/elasticsearch)
Vendor: Elasticsearch BV (Open Source)
Popular Use Cases:

- Amadeus IT Group, Facebook, Github, Netflix, Zalando SE,

Which problem can be solved: Index, search and provide analytics of a relatively large dataset
Brief description:

- Elasticsearch is a distributed, RESTful search and analytics engine
- It provides a distributed, full-text search engine with an HTTP web interface and schema-free JSON documents
- Elasticsearch is developed alongside a data-collection and log-parsing engine called Logstash, and an analytics and visualisation platform called Kibana
- Another feature is called "gateway" and handles the long-term persistence of the index

Alternatives: Apache Solr, Amazon CloudSearch

Data Processing: TensorFlow

Name: TensorFlow (www.tensorflow.org)
Vendor: Google Brain Team (Open Source)
Time in the market: ~2 years
Popular Use Cases:

- Clients include: Airbnb, DeepMind, Coca Cola, Google, ebay, Twitter
- Speech and Image recognition, Object tagging videos, Self-driving cars, Detection of flaws, Text summarization
- Mobile image and video processing, and Air, land, and sea drones

Brief description:

- A library for numerical computation using data flow graphs
- Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them
- The flexible architecture allows you to deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device with a single API

Alternatives: Apache SparkML and SparkMLib, Apache Flink

Technologies: Visualisation

- Kibana
- Tableau
- Plot.ly

Visualisation: Kibana

Name: Kibana (www.elastic.co/products/kibana)
Vendor: ElasticSearch (Open Source)

Popular Clients: Airbnb, BitBucket

Brief description:

- An open source data visualization plugin for Elasticsearch
- Visualization capabilities on top of the content indexed on an Elasticsearch cluster
- Put Geo Data on Any Map
- Time Series
- Analyze Relationships with Graph
- Explore Anomalies with Machine Learning (unsupervised ML using X-Pack)

Alternatives: Grafana

Visualisation: Tableau

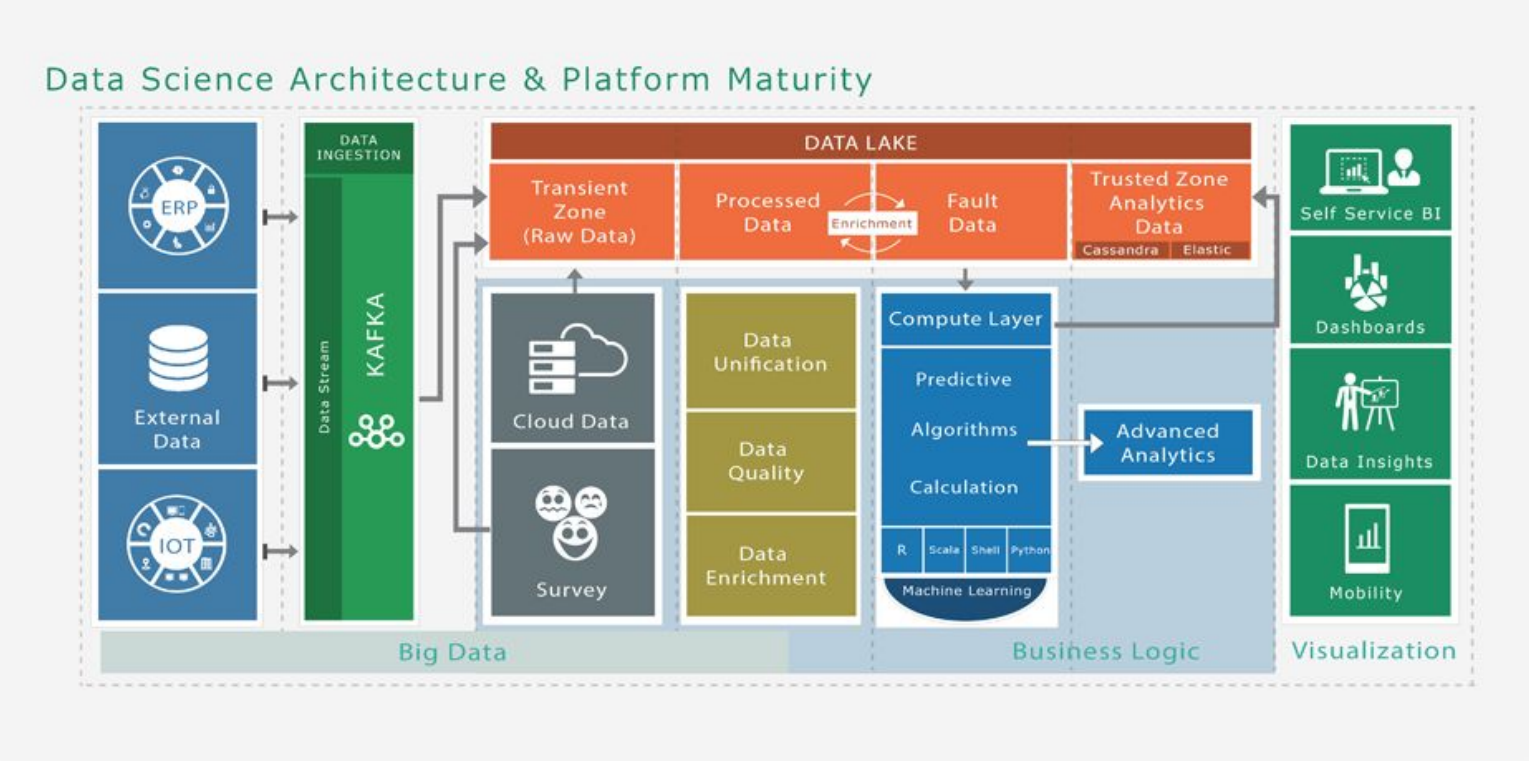
Name: Tableau (www.tableau.com)
Vendor: Tableau Software (Proprietary)
Time in the market: ~15 years
Popular Clients: Audi AG, LinkedIn, New York Times

Brief description:

- Tool specialized in visualization techniques for exploring and analyzing relational data

Alternatives: IBM Cognos, Qlik Sense Desktop, Microsoft Power BI, Tibco Spotfire

Recapping..



Big Data Landscape

I like it! What can I do next?

SQL for Data Scientists (next class)

Dig more about the technologies we discussed

Look for examples online, there're plenty

Questions?