An Overview Of Data Engineering Technologies

Abstract

Big Data is a buzzword and keeping track of the booming big data ecosystem is not easy. Even harder is to decide on a software stack, because comparison differs from use case to use case and most technologies are constantly refined.

This poster give an overview of common technologies for data engineering, categorizes them and lists their key features as a first step to an informed decision on an appropriate software stack.

Big Data Processing Frameworks

MapReduce processing engine

Can handle enormous datasets

5,476 ★

hadoop

YARN

A framework to manage computing resources in clusters.

HDFS

A distributed file system that provides high-throughput access to application data.

Resilient Distributed Datasets (RDD)

Streams as micro-batches

15,995 ★

APACHE Spark™

Best ML-Library

DAG scheduling

Lambda architecture

Kappa architecture

A distributed streaming dataflow engine

3,229 ★

Apache Flink

Entry-by-entry processing

Event time handling

End-to-end exactly-once guarantees

Tuple based

Hadoop Extensions

512 ★

Pig provides a high-level platform to write Hadoop Jobs.

1,680 ★

Adds data warehouse functionalities and a SQL-like abstraction.

OOZIE 351 ★

A workflow scheduler system for Hadoop jobs.

Data Access

1,728 ★

APACHE HBASE

4,147 ★

cassandra

- NoSQL
- Distributed database management system
- High Scalability
- Real time read and write access
- Based on Google's Bigtable.

Streaming Engine

7,289 ★

kafka

A distributed streaming platform. Publish and subscribe to a stream of records.

Notebook Platforms

3,515 ★

jupyter

3,370 ★

Apache Zeppelin

- Collaborative data analysis
- Literate programming
- Support for 20+ programming languages

Search Engines

1,383 ★

Solr

28,416 ★

elasticsearch

- Real time search in unstructured data
- Based on Apache Lucene
- Full text search
- Relevance scoring of results
- Faceted search
- Hit highlighting
- Rich document handling

Keys

★ Stars on Github

⋯➤ Supported by

Ringvorlesung "Data Engineering in der Praxis"

Adrian Ziegler

Bachelor of science

HPI Hasso Plattner Institut

Digital Engineering · Universität Potsdam