THE ZOO OF DATA ENGINEERING

A RANDOM WALK — PART I

CLOUDERA IMPALA

Impala is an open-source query engine connected to the Hadoop stack. It provides scalable parallel database logic, enabling SQL queries to data on HDFS and Hbase without moving the data.

GRAPHLAB

GraphLab is a machine learning framework for distributed systems with a focus on graph-based problems. As it has a rather small community and is not as integrated in the Hadoop ecosystem as e.g. Mahout is, it's full potential is not tapped yet.

BATCH PROCESSING

MYSQL

The classic relational database can be scaled on e.g. Hadoop clusters very well but as it needs very structured data the formatting process often leads to heavy performance issues on the cluster.

XING

DATA STORE

MAHOUT

Mahout is a machine learning framework that is optimized for cluster computing. It is widely used and has a strong community and documentation. On the other side it is not the most powerful ML library, e.g. GraphLab is up to x20 times faster for specific tasks.

KUDU

Kudu is an open-source storage engine working on top of the Hadoop stack and works column-oriented, hence relational. It combines fast writing with efficient columnar scanning for that real-time analytic tasks can be run on a single storage layer.

HBASE

Hbase is an open-source, distributed database and categorized as non-relational. Only one master controls all other slave nodes. It is not possible to run multiple masters in the same network which is why the master is a performance bottleneck especially when it faces a dropout.

BAKDATA · NEOFONIE

ABSTRACT

This poster shows a zoo with selected representatives of the (big) data engineering jungle. The animals are separated in different compounds so that only creatures of the same species live together. Every animal has a sign providing information about its characteristics and skills. I also tried to point to the problems we have discussed in lecture which come along with every technology animal.

So the viewer gets a general overview of data engineering technologies that are necessary to build a state-of-the art product. The shown solutions point to the underlaying problems which determine the discipline of data engineering and processing of data at big scales. Finally one gets an impression of problems that engineers face in the real world when companies use these technologies.

STUDENT

Moritz Hartmann
IT-Systems Engineering (Bachelor)

ADVISORS

Dr. Ralf Krestel,
Prof. Dr. Emmanuel Müller,
Prof. Dr. Felix Naumann,
Dr. Matthias Uflacker

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam