Summer Student Project Report

Dániel Stein (IT-DB-DBF)

2015 Summer

For 10 weeks —from 15th of June to 21st of August— I was a Summer Student at the CERN Database Group. This report summarizes my project and my progress.
Supervisors: Kacper Surdy, Zbigniew Baranowski. Project ID: 15465.

1 Task Description

Hadoop framework is becoming a top open source player on field of distributed system offering possibility of storing and processing big sets of data in a scalable manner. Nowadays it is also gaining momentum at CERN – there is an increasing interest in solutions based on Hadoop ecosystem in many areas including experiments, accelerator controls archives etc. It appears as a natural replacement of a traditional relational system whenever ad hoc analytical processing is a dominant part of data workloads.

Some of the systems at CERN already have offline replicas of archive data stored on a Hadoop created manually by a system administrator. Such process is very often time consuming and requires applying sequence of actions on the metadata and data itself.

The goal of the project is to automatize the process of data loading between Oracle database and Hadoop cluster by creation of a utility that interfaces with both systems via dedicated tools (Apache Sqoop) and applies all necessary actions in order to delivered ready-to-read data on Hadoop file system for high-end frameworks (like Impala or Spark). The tool should support incremental data loading on a time scope bases and should be configurable for each data set separately.

2 Challenges

Based on the task description and the initial consultations, my task was to build an automated, incremental Oracle–Hadoop data transfer framework and service with the following in mind.

It has to **control and monitor data transfers** using Sqoop, a CLI tool for bulk data transfer. Its aim is to **take the load off the system administrators** by executing the jobs with an automated tool. The system should **improve communication** with the users by notifying them about the status of the transfers.