Platform Interactions with Hadoop

Contents:

The Trifacta® platform interacts with your enterprise Hadoop cluster like any other Hadoop client. It utilizes existing HDFS interfaces and can optionally integrate with Hadoop components such as Spark and Hive for better connectivity and performance.

In standard deployment, the Trifacta platform reads files and stores results in HDFS. Optionally, it can execute distributed data processing jobs in Hadoop Spark.

The following diagrams illustrate how the Trifacta platform interact with Hadoop in various execution and deployment scenarios. Diagrams include the in-use client and server components and the ports over which they communicate.

- Ports are typically configurable, but the default ports are shown. See *System Ports*.

Run the Job Locally

- **T**he Trifacta application initially loads the head of the dataset using WebHDFS for display in the Transformer page.
- WebHDFS is also used to upload local files and browse HDFS during flow and dataset creation.
- The batch-job-runner component is triggered to gather a sample, which in turn reads data from Hadoop via WebHDFS.

WebHDFS requires access to HTTP port 50070 on the namenode server and a redirected request to HTTP port 50075 on any datanode containing queried data.

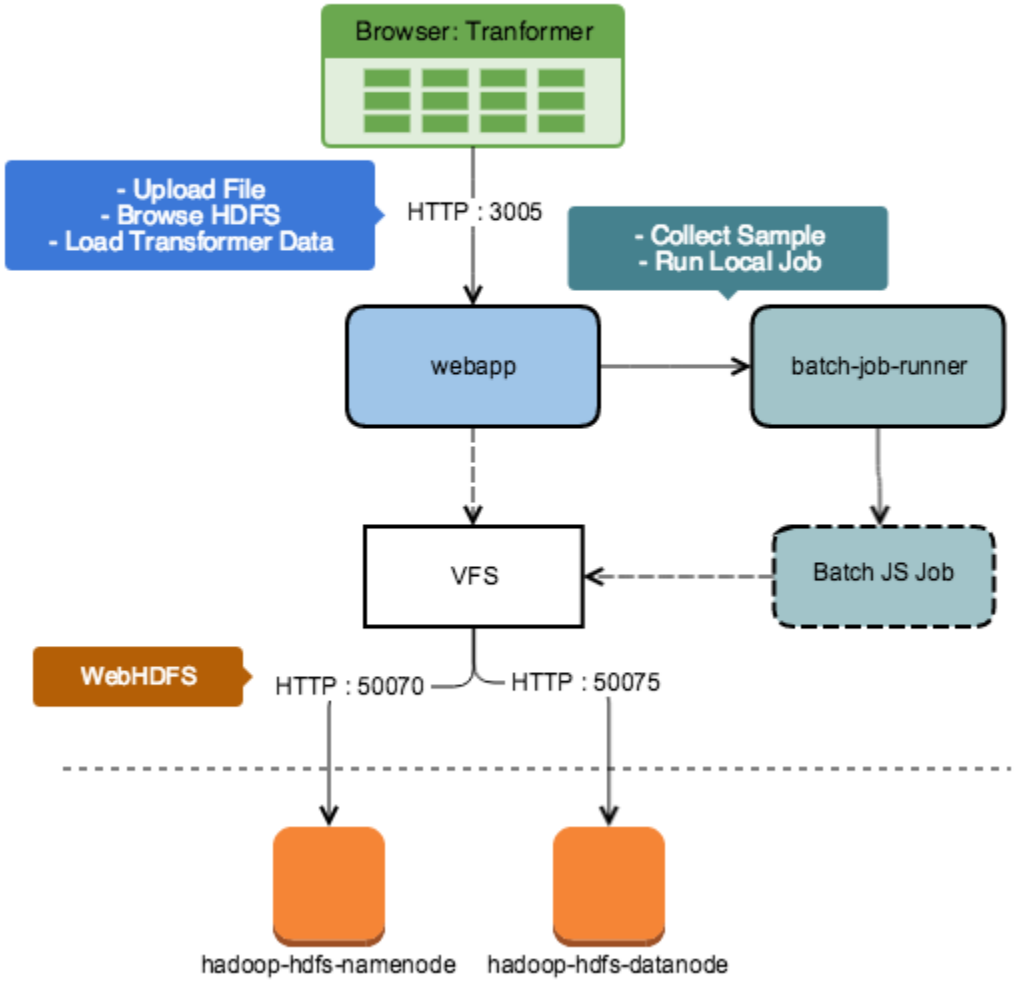

Figure: Run on Server

Run in Hadoop (YARN Cluster)

YARN jobs must distribute jobs via resourcemanager IPC port 8032. They also communicate with the Application Master created for that job on the cluster, which uses a new port within a configured range of ports.

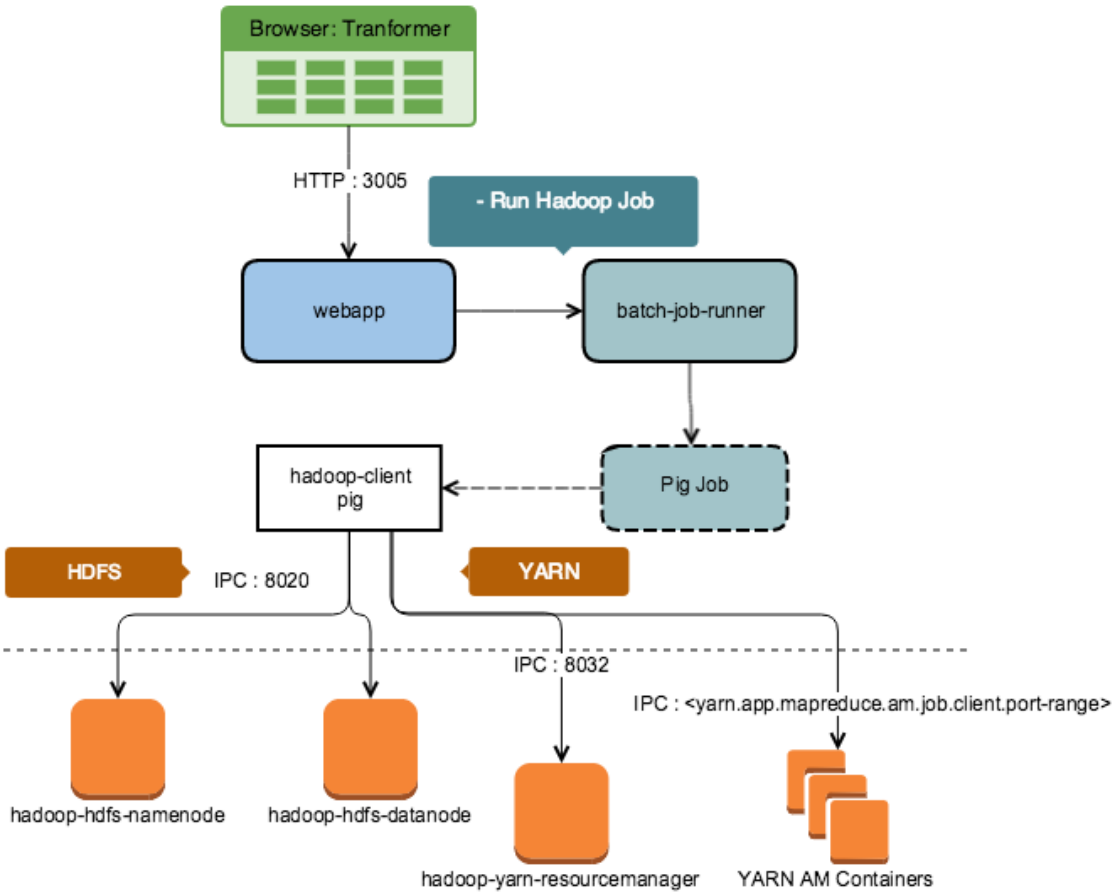

Figure: Run Job in Hadoop YARN Cluster

Run Job in Scala Spark

By default, the Trifacta platform executes transformation jobs and profiling jobs using the Scala version of Spark. This set of libraries can be deployed to nodes of the cluster from the Trifacta node, so that no cluster instance of Spark is required.

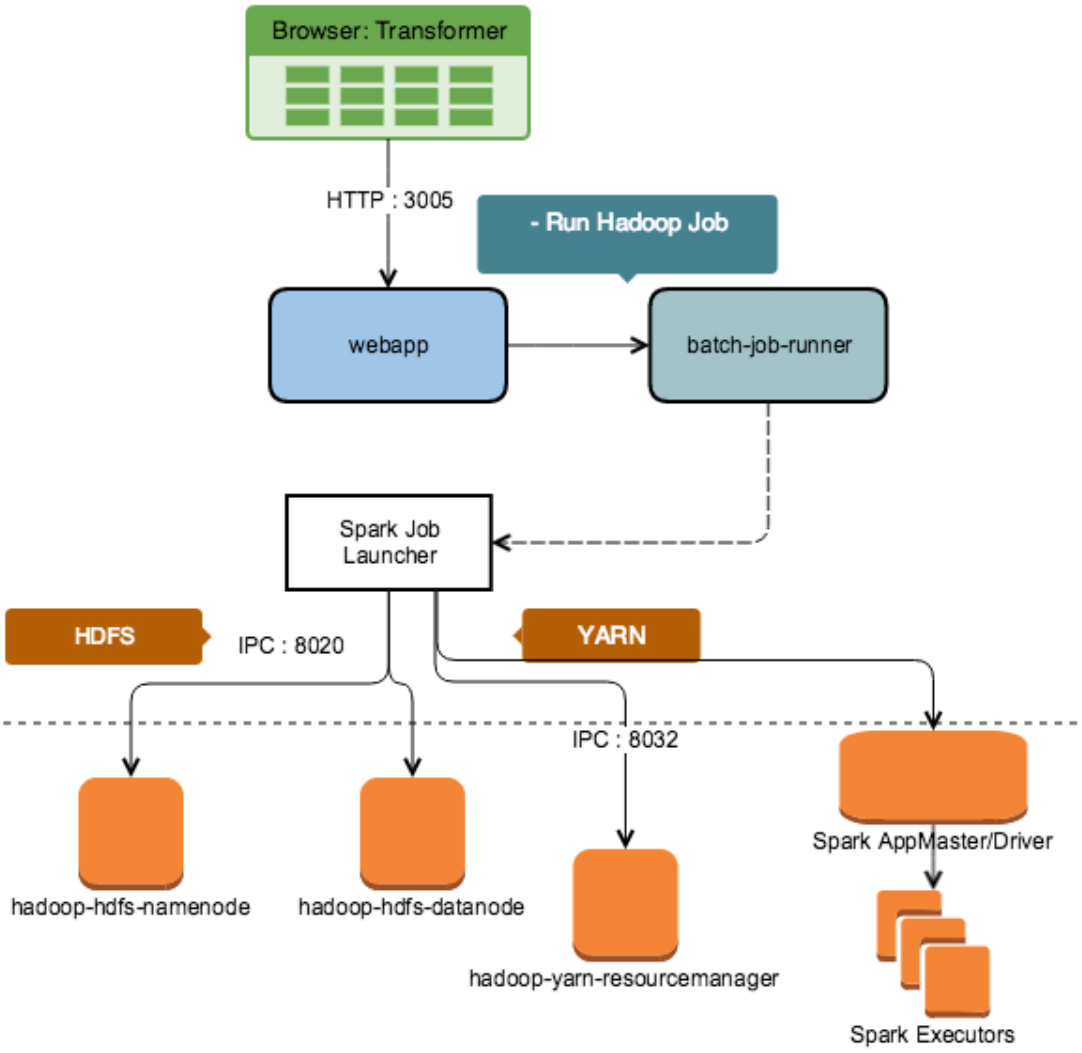

Figure: Run Job on Scala Spark

Run Profiling Job in Scala Spark

The following diagram shows the workflow for Scala Spark-based profiling jobs:

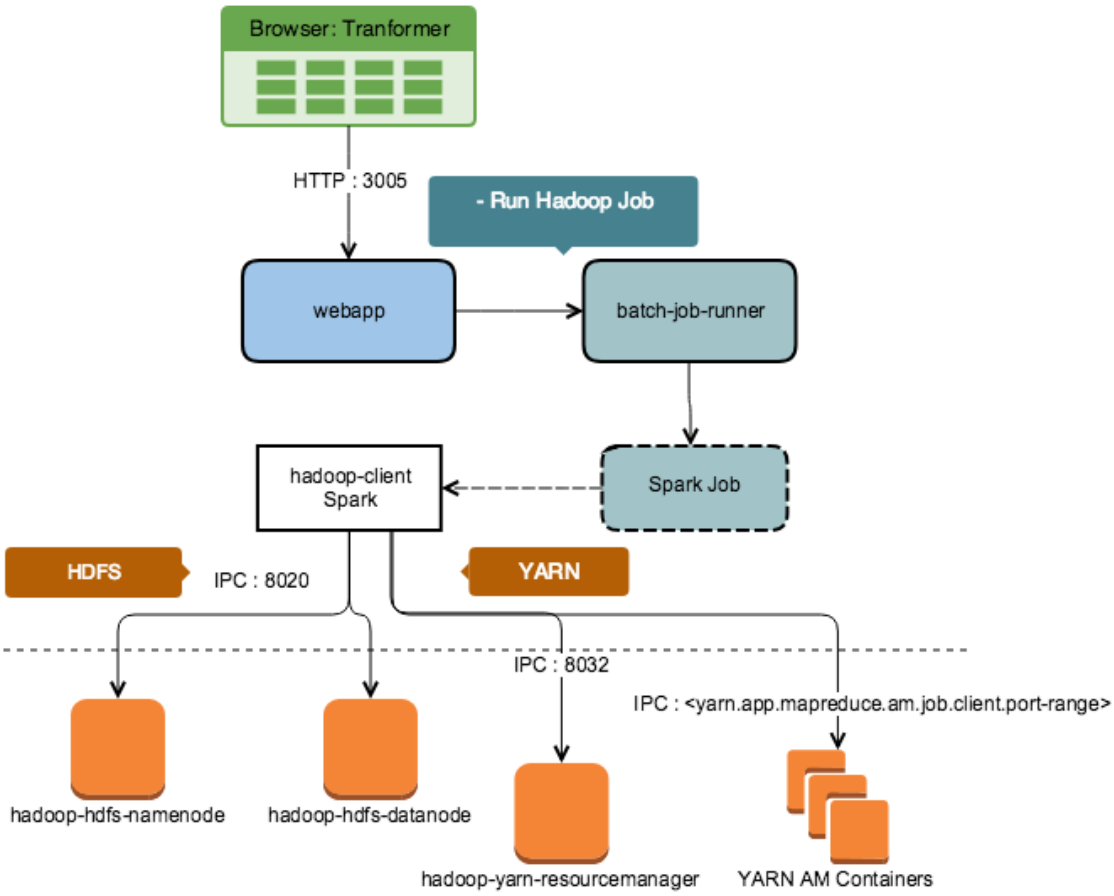

Figure: Run Profiling Job on Scala Spark

Coordination and Publishing Flows

The Trifacta platform uses Batch Job Runner, an Activitii-based service, for job orchestration. In the following publishing flow, the Trifacta platform is publishing results to Hive.

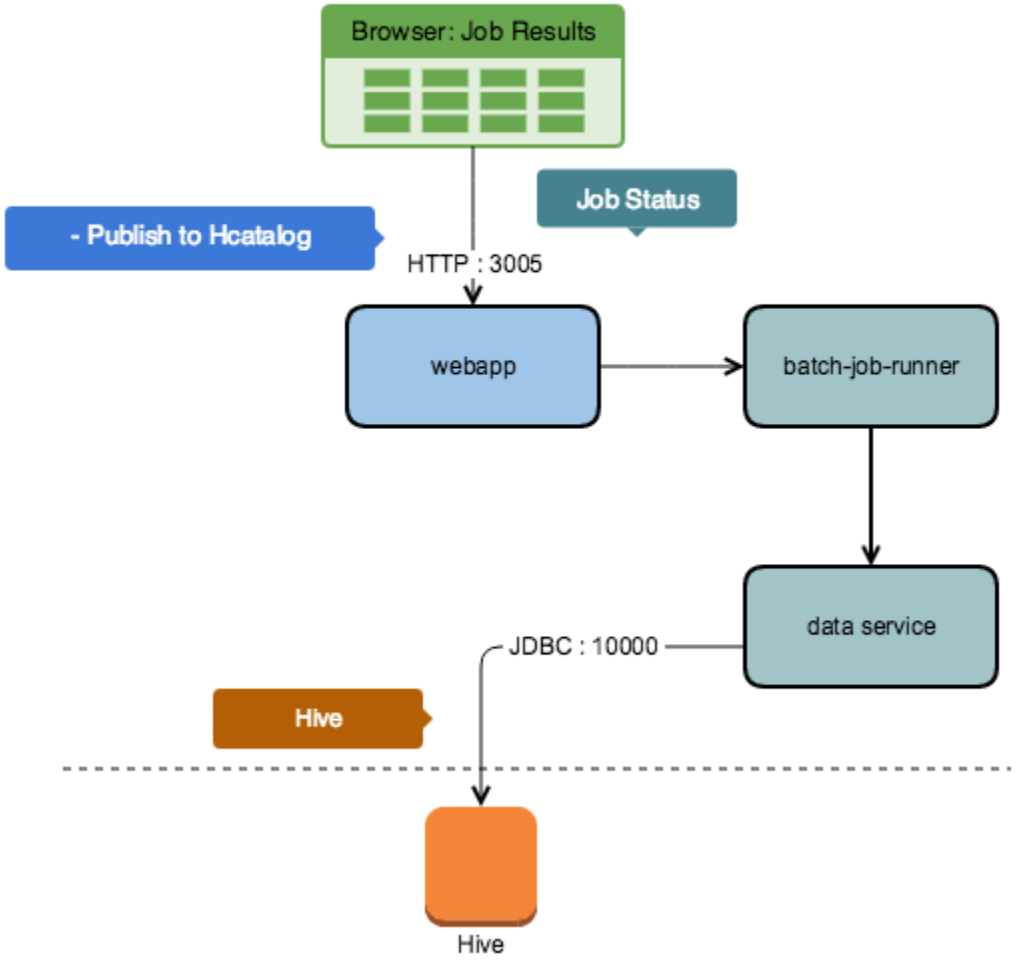

Figure: Coordination and publishing to Hive