# **Grupo Bimbo Demand Prediction - Final Report**

**By:**

**Tanya Malhotra (malhotra.t@husky.neu.edu)**

**Paritosh Arora (arora.par@husky.neu.edu)**

Contents

[**Groupo Bimbo Demand Prediction - Final Report** 1](#_Toc458440894)

[**Goals** 3](#_Toc458440895)

[**Data Source** 4](#_Toc458440896)

[**Approach** 5](#_Toc458440897)

[**Source Code** 6](#_Toc458440898)

[**Commands** 7](#_Toc458440899)

[**Screenshots** 8](#_Toc458440900)

[**Results** 11](#_Toc458440901)

[**Problems Faced** 11](#_Toc458440902)

[**References** 11](#_Toc458440903)

## **Goals**

1. **Business Goal:**

* Currently, the bakery’s demand is anticipated by the workers based on their personal experience. This needs to be changed by predicting the demand based on historical sales data in order to be more accurate.
* While doing so, Grupo Bimbo intends to reduce the amount spent on refunds to store owners with surplus product unfit for sale.

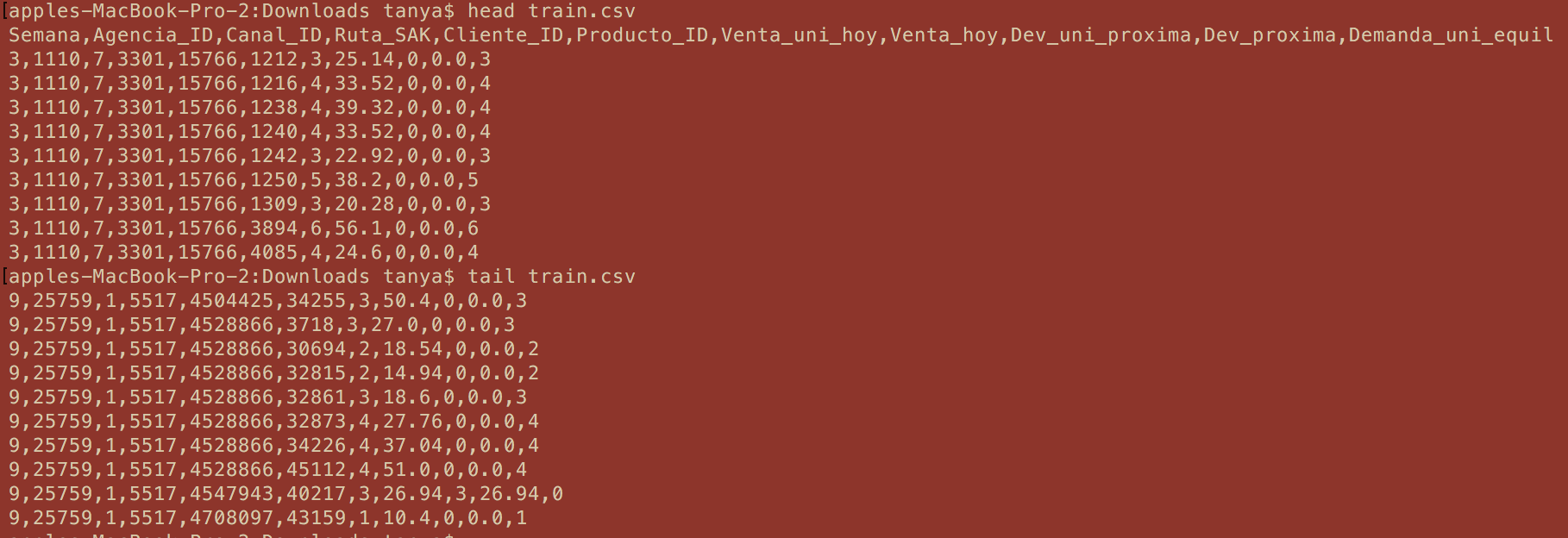
1. **Technical Goal:**

* Set up platform to analyze and process the big dataset
* Apply predictive analysis on provided dataset and evaluate the resultant model

## **Data Source**

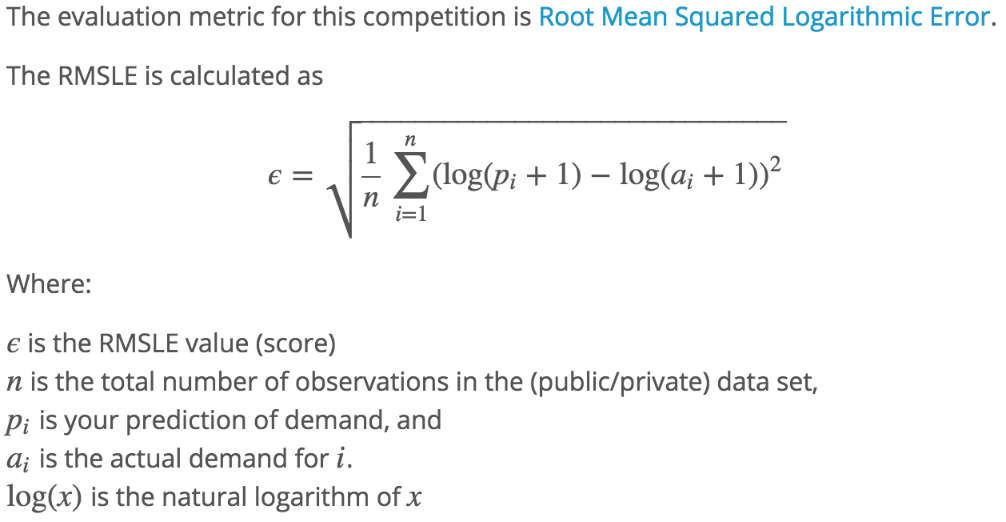
The data is sourced from one of the current Kaggle competitions:

<https://www.kaggle.com/c/grupo-bimbo-inventory-demand/data>

****

## **Approach**

1. **Parse & Transform Training Data:** The training data was initially the form of a csv file which was converted into ‘label’ and ‘features’ format which is the required input format for Spark’s machine learning API.
2. **Train Regression Model:** We chose Gradient Boosted Trees (GBT) regression algorithm to train our model for weeks 3-7 of the dataset (training data).
3. **Predict Demand Value:** Using the trained modelwe predicted the demand values for the weeks 8 and 9 (test data).
4. **Evaluate Model:** We calculated root mean square log error value using the predicted values and the associated actual values using the formula:

****

## **Source Code**

The source code is available at:

<https://bitbucket.org/arora_par/demandprediction>

Access to the above bitbucket repository has been given to following email IDs:

[vallabhaneni.l@husky.neu.edu](mailto:vallabhaneni.l@husky.neu.edu)

[c.konstantopoulos@neu.edu](mailto:c.konstantopoulos@neu.edu)

The readme file explains how to run the project.

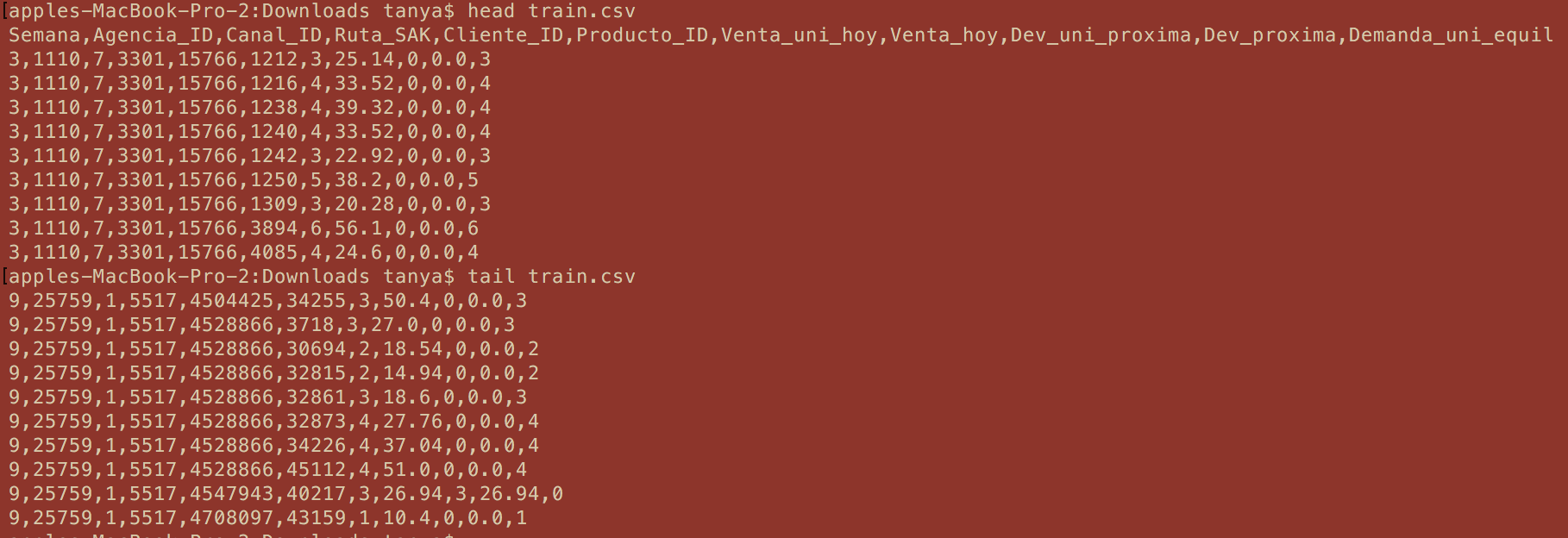
## **Commands**

* Compile Code
  + mvn clean compile package
* Copy jar to master node
  + scp -i bigdatakey.pem /Users/tanya/workspace/bakery/target/bakery-0.0.1-SNAPSHOT.jar [root@ec2-52-41-138-171.us-west-2.compute.amazonaws.com](mailto:root@ec2-52-41-138-171.us-west-2.compute.amazonaws.com):~/
* Move data to HDFS
  + scp -i bigdatakey.pem train.csv.zip [root@ec2-52-41-138-171.us-west-2.compute.amazonaws.com](mailto:root@ec2-52-41-138-171.us-west-2.compute.amazonaws.com):~/
  + ephemeral-storage/bin/hadoop dfs -mkdir input
  + ephemeral-storage/bin/hadoop dfs -put /root/train.csv.zip input
* Execute
  + spark/bin/spark-submit --class edu.neu.bigdata.course.BakeryDemandPrediction --master [spark://ec2-52-41-138-171.us-west-2.compute.amazonaws.com:7077](spark://ec2-52-41-138-171.us-west-2.compute.amazonaws.com:7077/) --packages com.databricks:spark-csv\_2.11:1.4.0 bakery-0.0.1-SNAPSHOT.jar

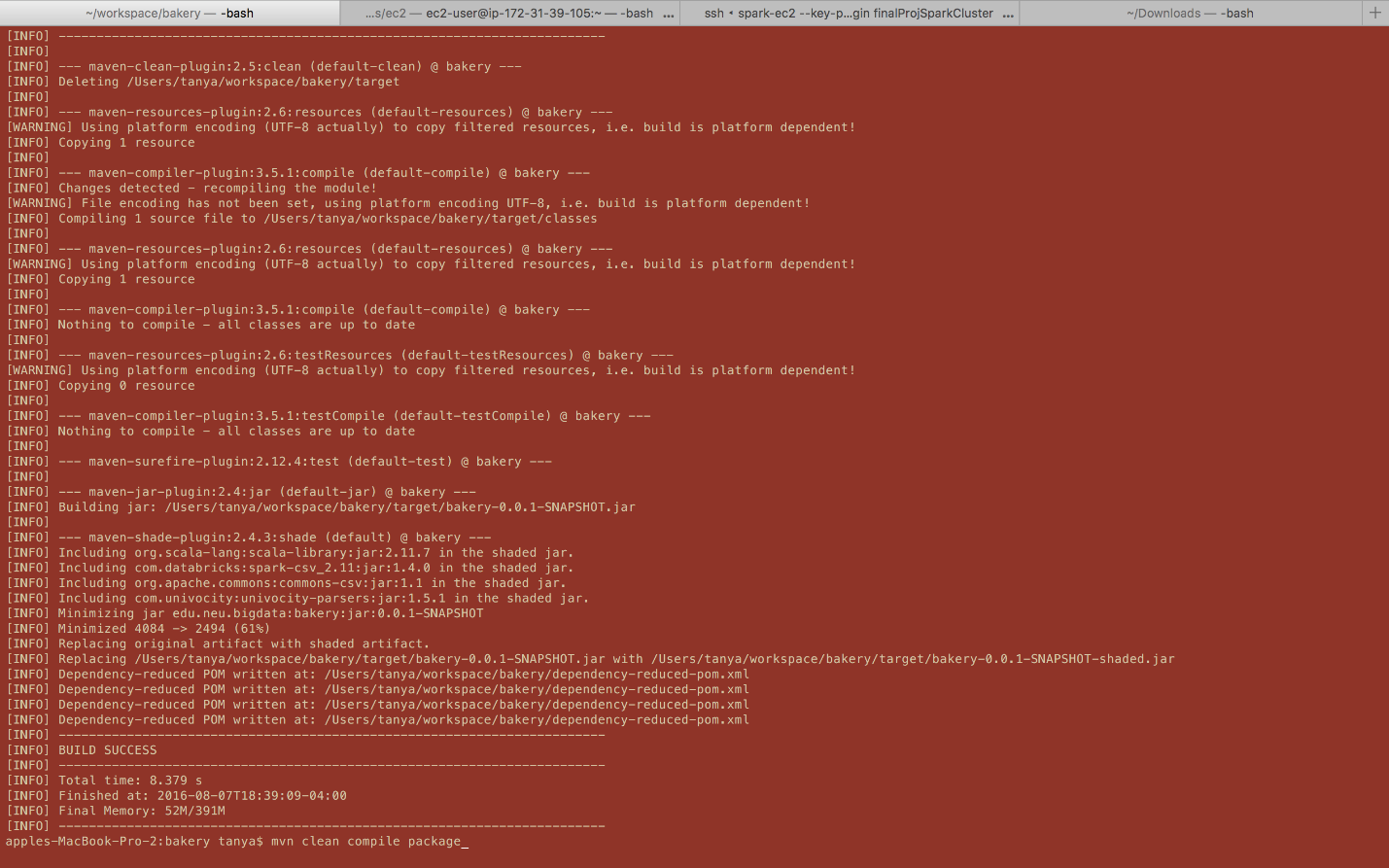
## **Screenshots**

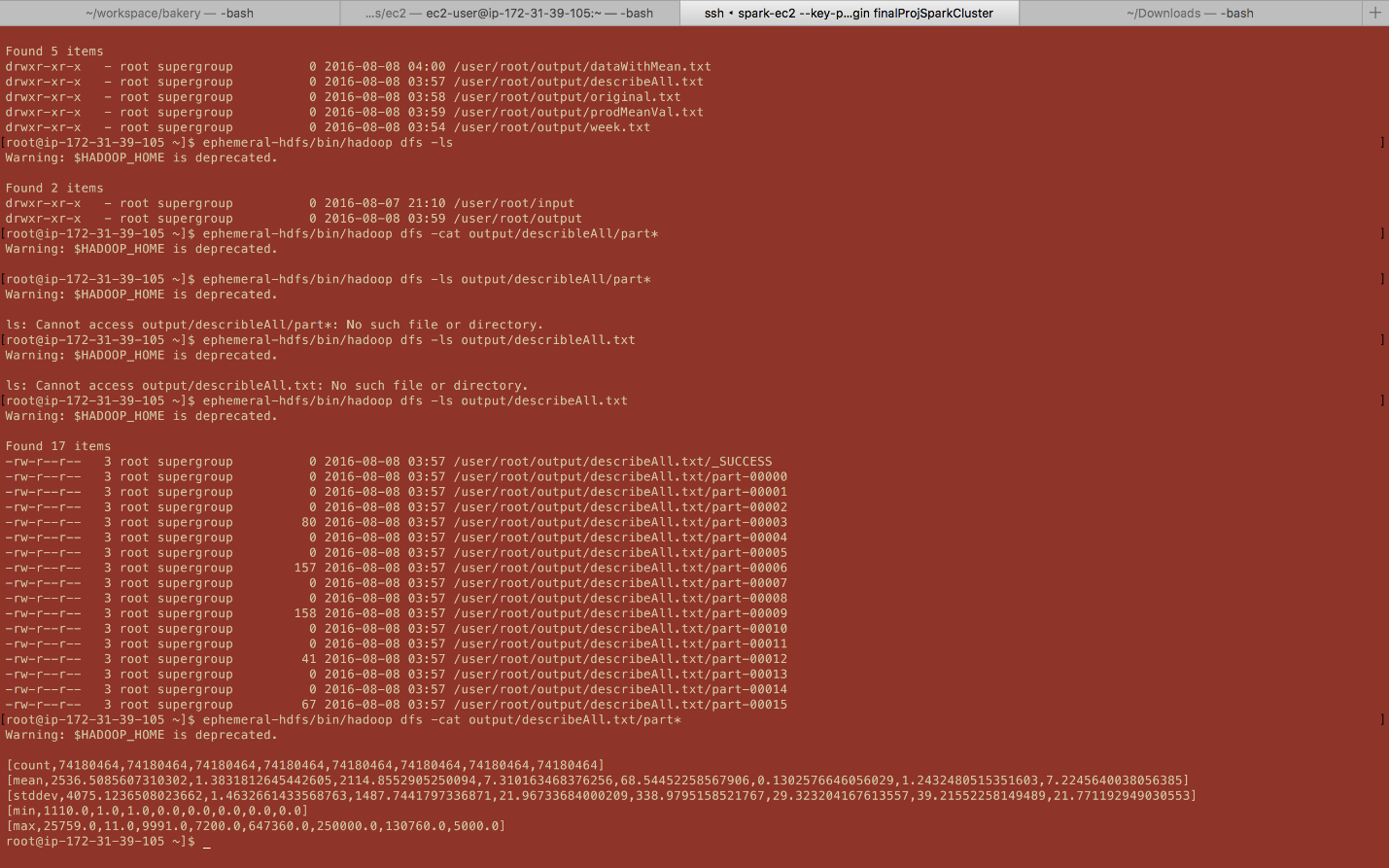
**Captured during various stages: Analysis, Build & Execution.**

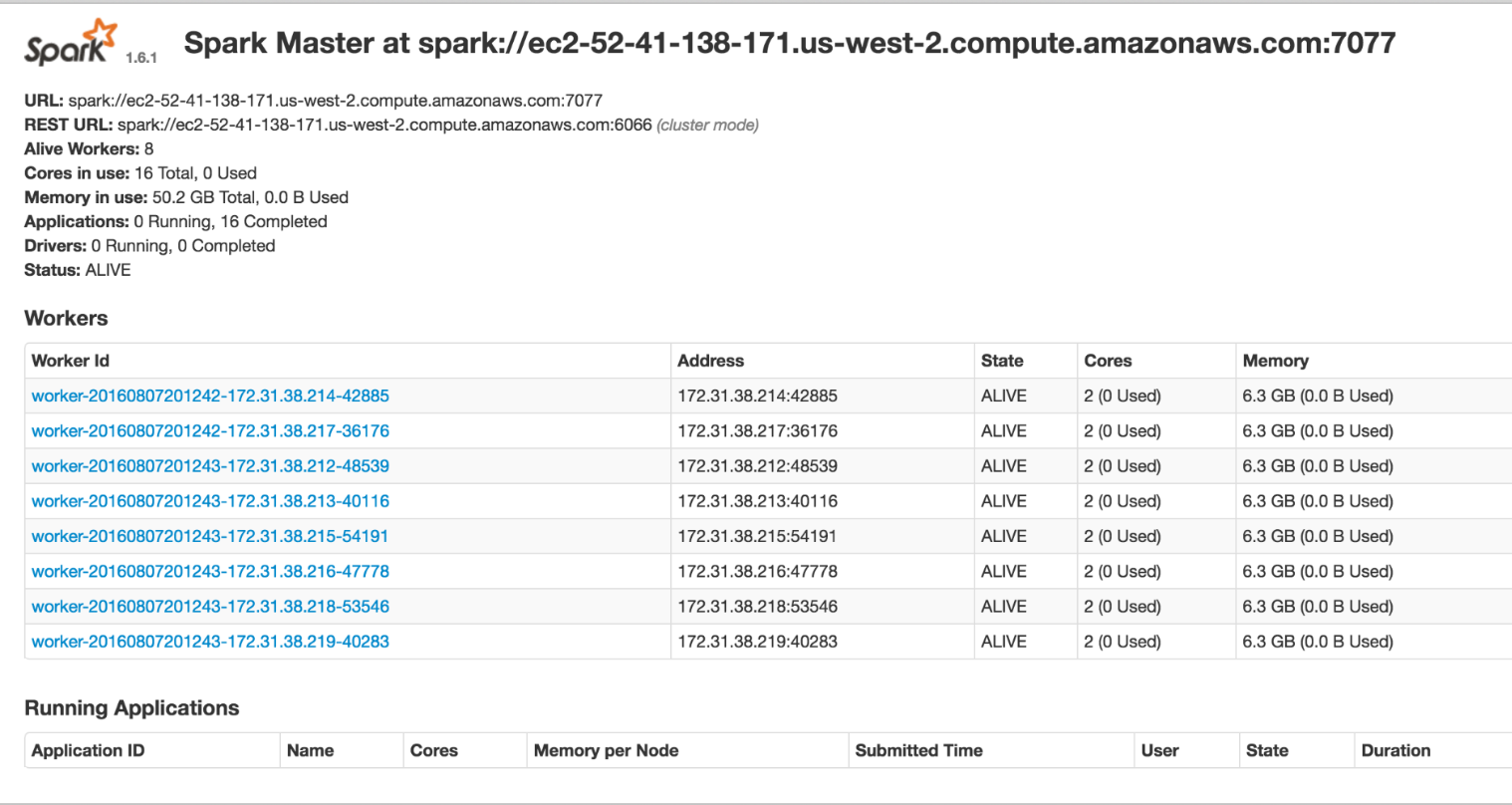
1. Initial analysis

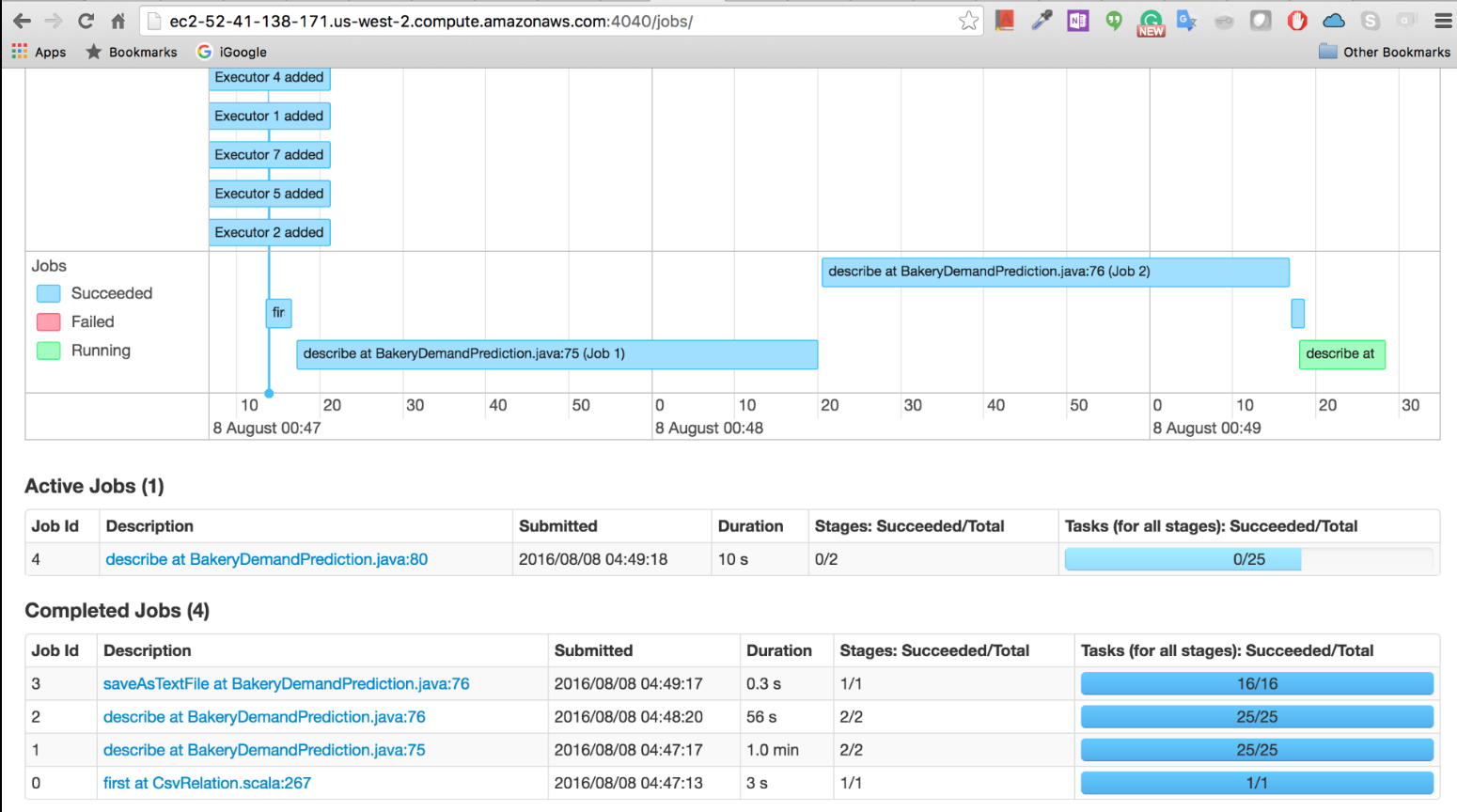
****

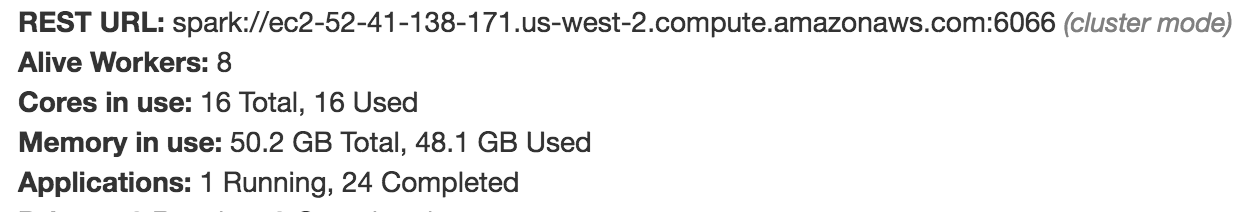
1. Maven build

****

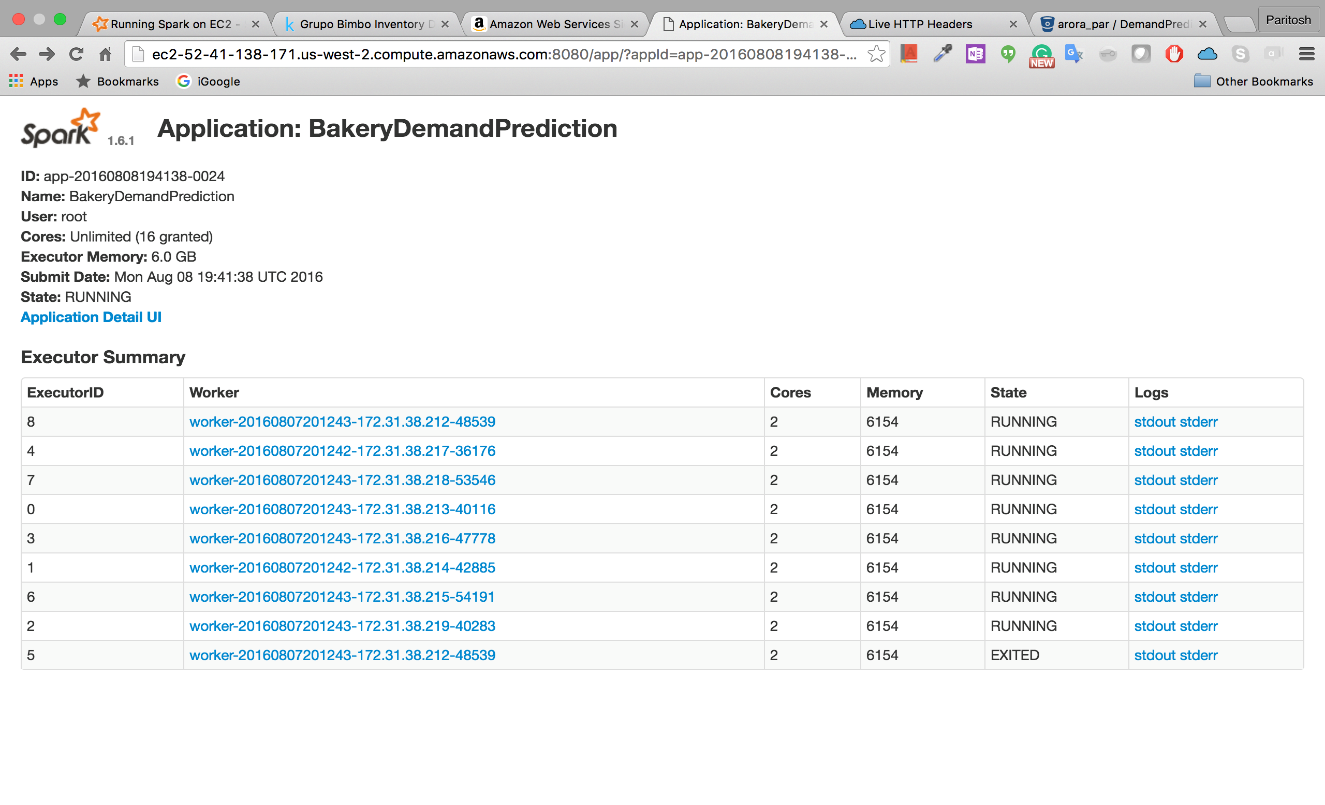
1. Describe DataFrame in Spark.****
2. Spark cluster with 8 Slave nodes.

****

1. Worker level monitoring details.
2. Maximum utilization of available CPU cores and RAM in the cluster.

****

1. Workers working on current application



## **Results**

We computed the root mean square log error to evaluate the accuracy of our model. The value of 0.3102 is good enough to be regarded as a successful model.

## **Problems Faced**

* Size of the dataset – Too big to run on standalone system

**Solution**: Set up cluster on AWS.

* Cluster setup – Access issues for command line tools

**Solution**: Set up IAM users and access policies.

* JDK compatibility issue – 8 vs 7 on localhost and cluster

**Solution**: Use Java 7 on both the platforms – no more lambda expressions.

* Spark compatibility issues – 2.0.0 vs 1.6.2

**Solution**: Used version 1.6.2

## **References**

<https://www.kaggle.com/c/grupo-bimbo-inventory-demand>

<http://spark.apache.org/docs/1.6.2/ml-classification-regression.html#gradient-boosted-tree-regression>

<http://spark.apache.org/docs/1.6.2/api/java/index.html>

<http://spark.apache.org/docs/1.6.2/ec2-scripts.html>