ACADGILD

BANKING DATA ANALYSIS USING HADOOP

About ACADGILD

ACADGILD is a technology education startup that aims to create an ecosystem for skill development in which people can learn from mentors and from each other.

We believe that software development requires highly specialized skills that are best learned with guidance from experienced practitioners. Online videos or classroom formats are poor substitutes for building real projects with help from a dedicated mentor. Our mission is to teach hands-on, job-ready software programming skills, globally, in small batches of 8 to 10 students, using industry experts.

ACADGILD offers courses in

Enroll in our programming course
& Boost your career

- ANDROID DEVELOPMENT
- DIGITAL MARKETING
- MACHINE LEARNING WITH R
- BIG DATA ANALYSIS
- JAVA FOR FRESHER
- BIG DATA & HADOOP ADMINISTRATION
- FULL STACK WEB DEVELOPMENT
- NODE JS
- CLOUD COMPUTING
- FRONT END DEVELOPMENT (WITH ANGULARJS)

Watch this short video to know more about ACADGILD.

Disclaimer

This material is intended only for the learners and is not intended for any commercial purpose. If you are not the intended recipient, then you should not distribute or copy this material. Please notify the sender immediately or click here to contact us.

Published by
ACADGILD,
support@acadgild.com

Table of Contents

Big Bank, Big Problem

A leading banking and credit card services provider is trying to use Hadoop technologies to handle and analyze large amounts of data.

Currently, the organization has data in the RDBMS but wants to use the Hadoop ecosystem for storage, archival, and analysis of large amounts of data.

Creating the Hadoop Cluster and Deploying Test Codes

For any Big Data project, a proper environmental setup is mandatory. In real world scenarios, a Hadoop cluster can consist of 40K clusters/nodes. As it is not possible for us to build such a huge setup in an academic lab, we will be going with minimal 2 node cluster and simulate the ingestion, cleaning, analysis and send the report to the downstream team.

Create a Pseudo-Distributed Node Cluster and deploy all the test codes using Maven and SBT.

Data Ingestion

Bring data from RDBMS to HDFS. This data import must be incremental and should happen every 2 minutes.

Following tables have to be imported:

Table	Incremental Column
loan_info	Loan_id
credit_card_info	Cc_number
shares_info	Gmt_timestamp

All these data must be encrypted in HDFS. The HDFS data should be compressed to store less volume.

The Sqoop password must also be encrypted.

Table Details

loan_info

This table has following columns:

Loan_id	Loan Identifier
User_id	User Identifier
last_payment_date	Date on which the last payment was made
payment_installation	Amount payable in installations
Date_payable	Date in month when payment is made by the user

credit_card_info

This table has the following columns:

cc_number	Credit card number
user_id	User Identifier
maximum_credit	Maximum credit allowed by the user
outstanding_balance	Current outstanding balance by the user
due_date	Due date for the credit card payment

Shares_info

This table has the following information:

Share_id	Share Identifier
Company_name	Name of the organization
Gmt_timestamp	Timestamp of the share price
Share_price	Value of share

Analysis

- Find out the list of users who have at least 2 loan installments pending.
- Find the list of users who have a healthy credit card but outstanding loan account.

Healthy credit card means no outstanding balance.

- For every share and for every date, find the maximum profit one could have made on the share. Bear in mind that a share purchase must be before share sell and if share prices fall throughout the day, maximum possible profit may be negative.

Archival

The organization has a lot of survey data scattered across different files in a directory in local file system. Provide a mechanism to effectively store the small files in Hadoop. It is expected to pack small files together before actually storing them in HDFS.

Survey files have the following structure:

survey_date	Date on which the survey was conducted
survey_question	Questions asked in the survey
Rating	Ratings received (1 - 5)
user_id	User ID who responded
survey_id	Survey ID

The following analysis is expected from survey files:

The following analysis is expected from survey files:

- How many surveys got an average rating of less than 3, provided at least 10 distinct users gave the rating?
- Find the details of the survey which received the minimum rating. The condition is that the survey must have been rated by at least 20 users.

The organization also has lots of e-mails stored in small files.

The metadata about the e-mail is present in an XML file email_metadata.xml

Read the XML file for the e-mail structure and pack all the e-mail files in HDFS.

Following analysis is expected from the e-mail data:

- Which is the longest running e-mail?
- Find out the list of e-mails that went unanswered.

ACADGILD

✉ support@acadgild.com | www.acadgild.com | ☎ 8880025025