EXPLORE | DIGITAL SKILLS

Processing Big Data
Predict Overview

Predict contents

EXPLORE | DIGITAL SKILLS

Predict overview

Prove your ability to master data engineering-related skills as you build an ETL process.

In this predict, you're a Data Engineer tasked with **ingesting** and **transforming** stock market data into a readable, reliable, and robust format. This stock market data contains historical daily prices for all tickers currently trading on Nasdaq.

To accomplish this, you will need to design and implement the **extract** and **transform** phases of an ETL. However, since stock market data is generated daily, these ETL phases are **time-sensitive** and needs to be **scalable**.

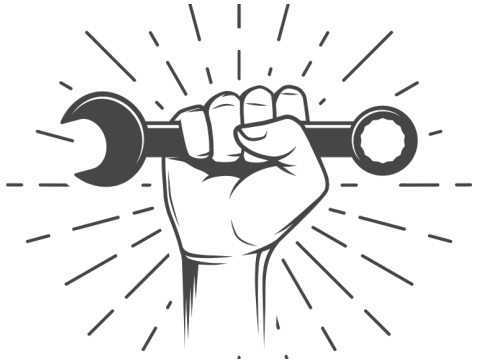
As a result, your tasks will include using **Apache Spark** to harness parallel processing in the transformation phase, profiling the data set according to the **six dimensions of data quality**, and implementing data quality testing with **Deequ**.

Predict setting: Transforming big data

Complete three tasks to build the **extract** and **transform** phases of an ETL for batch data.

The ingestion and processing of stock market data in a time-sensitive and scalable manner will provide functional datasets for stock market models and analyses. Having readable, reliable, and robust datasets enable data scientists to more easily use this data, with the certainty that it is correct.

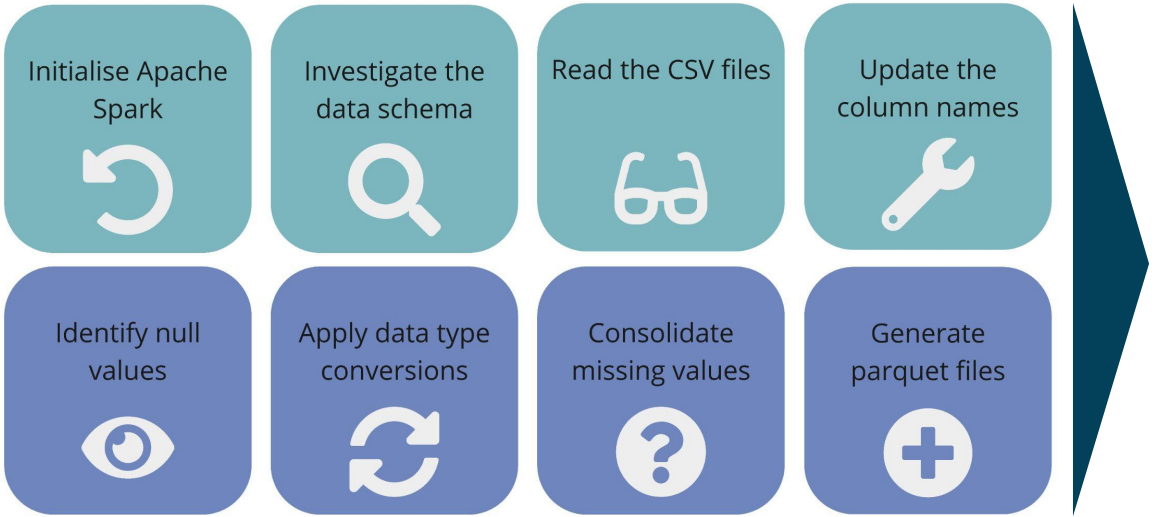
Task 1	Design and implement the data ingestion process and perform elementary data cleaning transformations by using Apache Spark.
Task 2	Design and implement transformations for dataset exploration and profiling according to the six dimensions of data quality.
Task 3	Design and implement data quality tests for a dataset using the Apache Spark library, Deequ.



EXPLORE | DIGITAL SKILLS

Part-I: Data ingestion

Design and implement the data ingestion process and perform elementary data cleaning on a dataset.



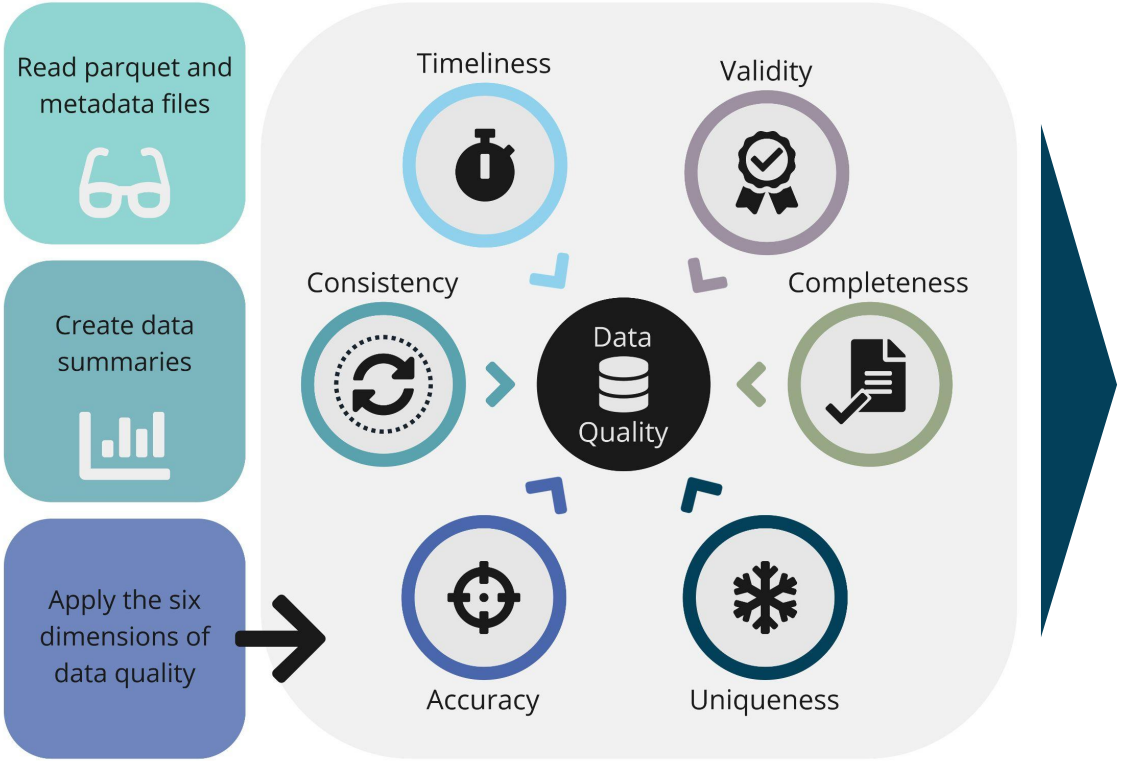
The data ingestion and cleaning steps within Part-I of this predict.

Learning outcomes

- Implement the first step of ETL, namely Extract:
  - Use Pandas and Apache Spark appropriately to read CSV files;
  - Apply common data engineering transformations, such as extraction, parsing, translation, filtering, and imputation.

Part-II: Dataset Profiling

Design and implement transformations for dataset exploration and profiling according to the six dimensions of data quality.



The data profiling steps, including the six dimensions of data quality.

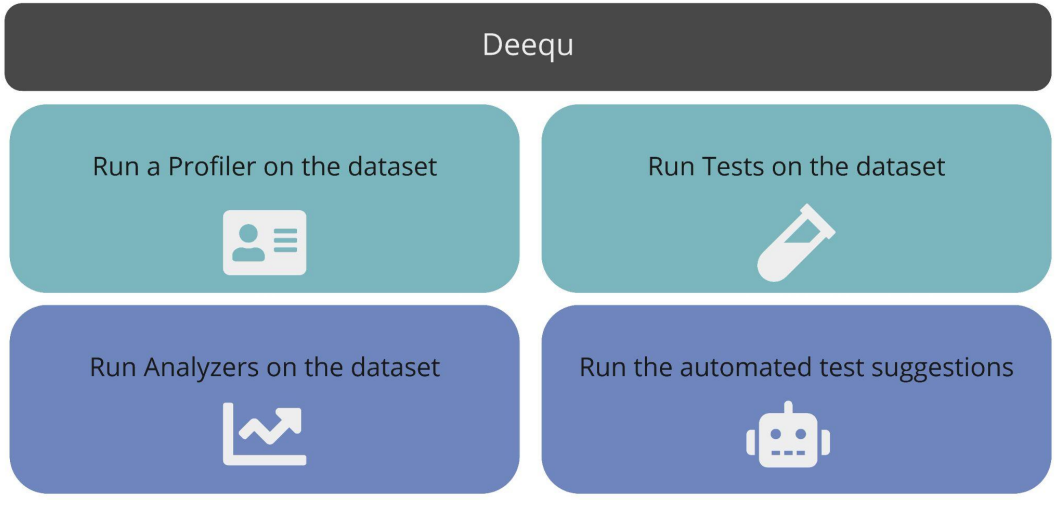
Learning outcomes

- Implement data transformations to prepare readable, reliable, and robust datasets:
  - Use transformations such as data filtering, summarisation, imputation, and enrichment.
- Understand and implement the six dimensions of data quality:
  - Use histograms to analyse the data distribution;
  - Use data transformations to determine the accuracy, consistency, timeliness, validity, completeness, and uniqueness of a dataset.

EXPLORE | DIGITAL SKILLS

Part-III: Dataset Testing

Design and implement data quality tests for a dataset.



The various Deequ classes and methods to be implemented for data quality testing.

Learning outcomes

- Design and implement data quality tests:
  - Use Deequ to profile, analyse, and test a dataset;
  - Specify metrics and configurations to test data quality.

EXPLORE DIGITAL SKILLS

Your mission: Complete the tasks, be an awesome data engineer

Demonstrate your knowledge of big data processing by smashing out the three tasks you've been challenged with!

Predict instructions

Processing Big Data Predict

To complete the predict successfully, you'll need to perform all the steps associated with each task. To help guide you, we've provided a Markdown that describes:

- The predict setting and the requirements of the transformation processes;
- The task steps for the various transformation required upon data ingestion;
- A set of high-level task steps to assist in profiling the dataset; and
- A set of high-level task steps to assist in designing and implementing data quality tests.

⏱️ **Assessment Notification** ⏱️

At the end of *each* task, **you'll be required to complete an MCQ** that will assess your knowledge of the content covered within the predict steps. As such, your goal shouldn't be to merely complete the tasks, but instead should be to understand the technologies and techniques covered therein.

Next steps

The following points should guide you around salient aspects for completing the predict.

Instructions

Go to this link and follow the instructions in the README.md file to get started with your tasks.

Need help?

You'll be expected to show increasing autonomy throughout this predict. To come to a solution, you'll need to make use of Apache Spark documentation, the awslabs/deequ GitHub repo, and some relevant tutorials.

Please consult the student forum for Processing Big Data predict-related FAQs.

Submission

This predict will be graded automatically. To ensure that you are assessed fairly, critically evaluate if your solutions produce the desired end result at each ***Predict Task*** step.

Your **predict deadline** will be communicated to you within the **Predict tab on Athena**.

EXPLORE | DIGITAL SKILLS