**PYSPARK AND SPARKCODING CHALLENGE**

**Name:** S.Nikhil Sai

**Date:** 26-11-2024

1. **ETL with Pyspark:**

ETL is a process in data engineering where data is Extracted from various sources, Transformed to make it usable, and Loaded into a destination, typically a data warehouse or database.

* **Extract:** Extracting the data from different sources
* **Transform:** The raw data is cleaned, filtered, enriched, and reshaped.
* **Load:** The processed data is saved back to a destination, such as a data lake, a distributed database, or a cloud storage service.

**Advantages of using ETL with pyspark:**

* PySpark processes massive amounts of data across distributed clusters, making it ideal for large-scale ETL operations.
* Provodes programmatic ETL, i.e., ETL operatins could be performed in coding commands using py spark
* It optimizes resource usage in a distributed cluster, reducing costs associated with large-scale ETL operations.
* PySpark can integrate with popular big data ecosystems like Hadoop, Hive, and HDFS.
* PySpark's Python API allows developers to write clean, reusable, and maintainable code

1. **Pyspark and spark SQL operations:**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**