CDS
Capital Data services

Technical Architecture Overview

CDS is very broad and contains of various parts: ingestion, change, business computations and reporting parts. The following is an outline of the CDS which is liable for source data ingestion and essential transformation (preparation of final views).

There are several types of data sources, the main one is a flat file, but there is also a possibility to source with direct SQL queries. For ingestion Nifi Service is used. This is a service for which data flow can be set up and where validation, file name comparison, dataset initialization and many other important things happen.

A stream for transferring flat files is designated "Cluster Data Ingestion" where transferring RDBMS data is refered as "Generic Data Ingestion" . For the stream, Nifi awaits a file begin put in the ingestion folder. The consequences of any stream are a parquet file put in the hdfs framework i.e hadoop cluster and Hive tables and a record in data catalogue service. Along these pipelines, the ingestion part is finished and the data can be tracked down in Hive and Impala in separate tables.

This way, we have the same data in Hadoop structures, as we have it in a source file. It is not transformed in any way, but new columns (such as version, date, etc.) can be appended. Before we can provide the data for business team to use, we need to transform it in a specific way (e.g. multiply market business prices with fixed USD rates). This kind of transformation is done in pyspark data processing service. To trigger spark jobs, we need : a spark job application details in job. A spark job application details is set directly in dependency management service. Then, dependency management service need a meta data entry to call a transformation.

A meta data entry is a data record put in the data catalog. Some datasets can be triggering and will call an actual transformation pyspark job, some of them don't need to trigger any job and used only outside part of CDS work. The results this process again are a parquet file put in the hdfs file system or in hadoop cluster and a record in data catalog service.

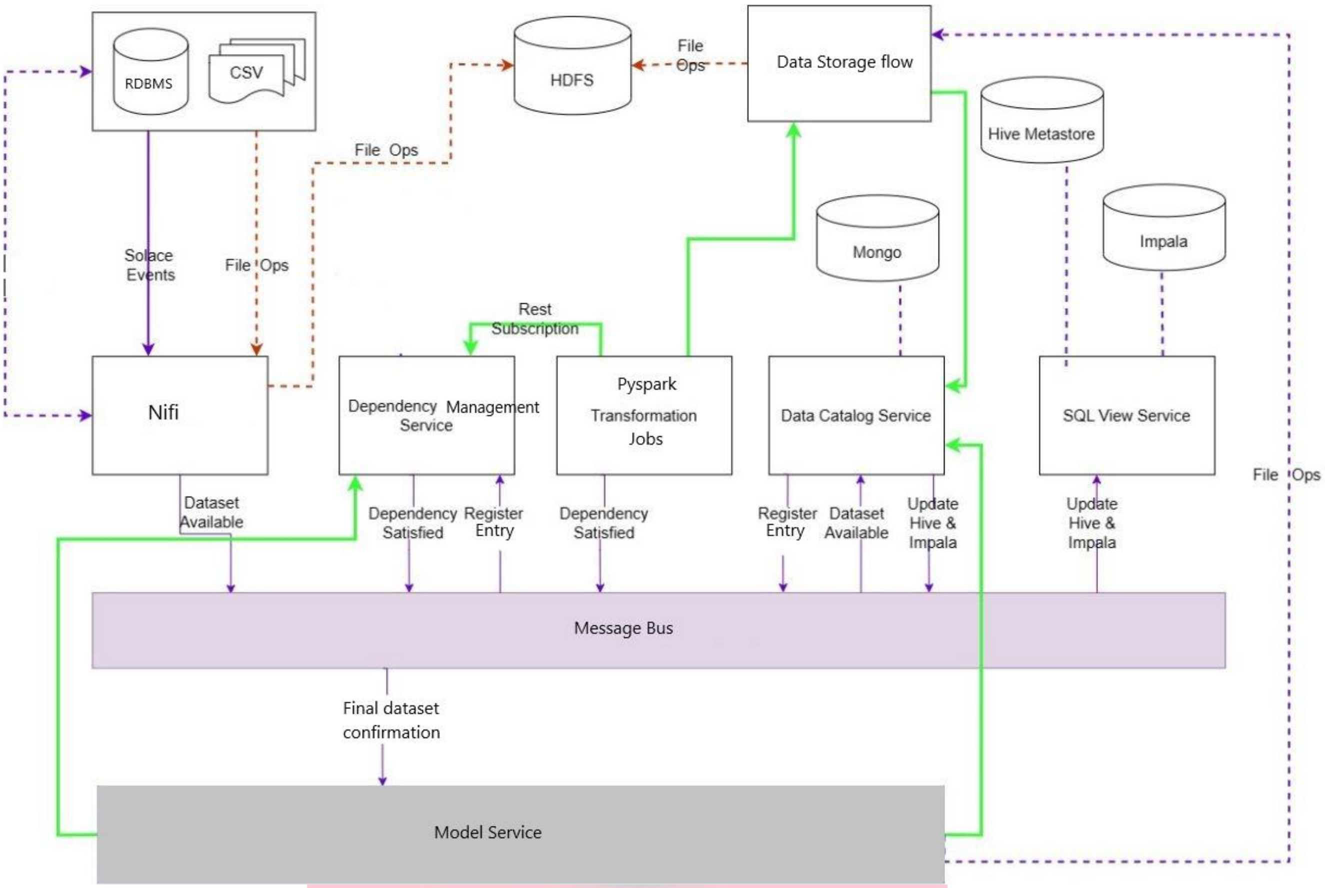
There is a new dataset table related with results of transformation process of modification and it will add new attributes on existing data tables. After transformation pyspark jobs completed, the resultant dataframe can be found in Hive and Impala in respective tables. Such tables have their names starting in upper case and is called "Business views". Dependency Manager Service and Data Catalog Service's data is kept in Mango database.

In the event that one of administrations service isn't accessible, the messages go to dead message line. At the point when transformationed dataset is prepared, it is consumed by Model service for additional business change and planning of information for further reports. Is additionally utilized for some reporting team groups for gathering multiple reports.

Purpose/Use Cases Json Meta Data available in data catalogue service is an summary of data stored in Hive tables or in HDFS path. Below are the key use cases that the Json Meta Data provide to the CDS system:

- Provides list of tables available in CDS (i.e. raw tables, business view tables after transformation and model tables)
- Register a datasets that is available in CDS that can be accessed by other services like Model service.
- Provide metadata information for a dataset which includes HDFS file location, input datasets and validation (like rowcount validation, file validation)
- Provide the "view" of each dataset used by CDS that will be used by other component when reading data via PYSPARK
- Used to apply and check input datasets for CDS Pyspark transformations .

Architecture Diagram


RDBMS
CSV
HDFS
File Ops
Data Storage flow
File Ops
Hive Metastore
Impala
Mongo
Solace Events
File Ops
Rest Subscription
Nifi
Dependency Management Service
Pyspark Transformation Jobs
Data Catalog Service
SQL View Service
Dataset Available
Dependency Satisfied
Register Entry
Dependency Satisfied
Register Entry
Dataset Available
Update Hive & Impala
Update Hive & Impala
File Ops
Message Bus
Final dataset confirmation
Model Service