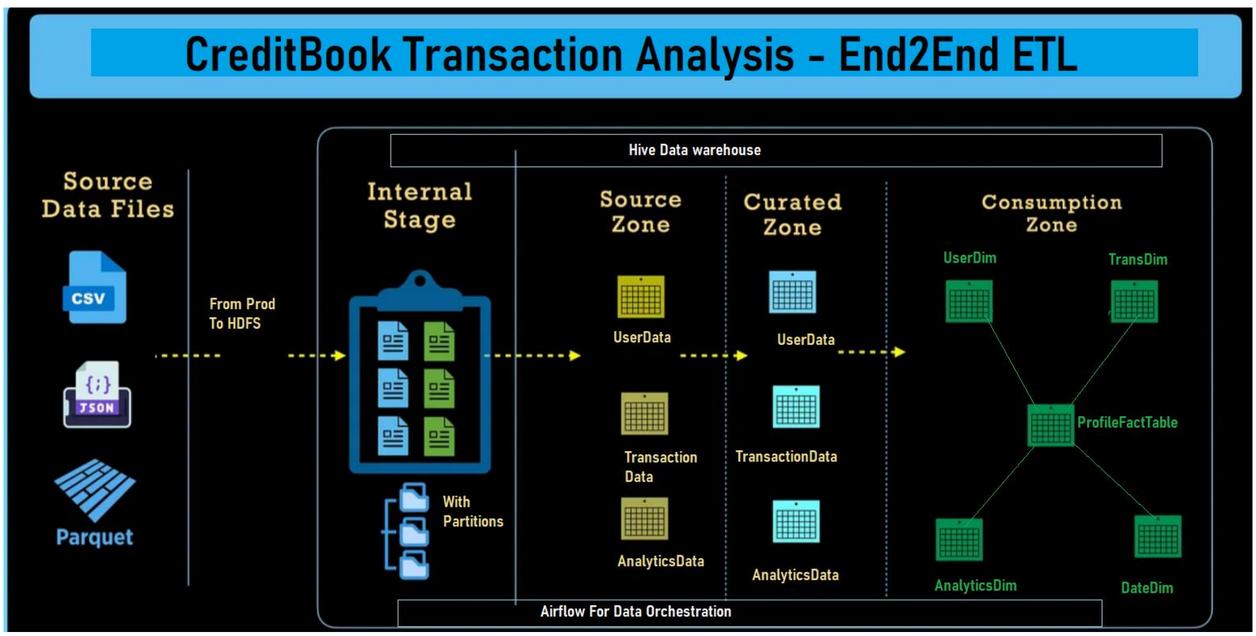


END2END ETL PIPELINE

ABSTRACT

Use case study For The Credit Book

Muhammad Affan

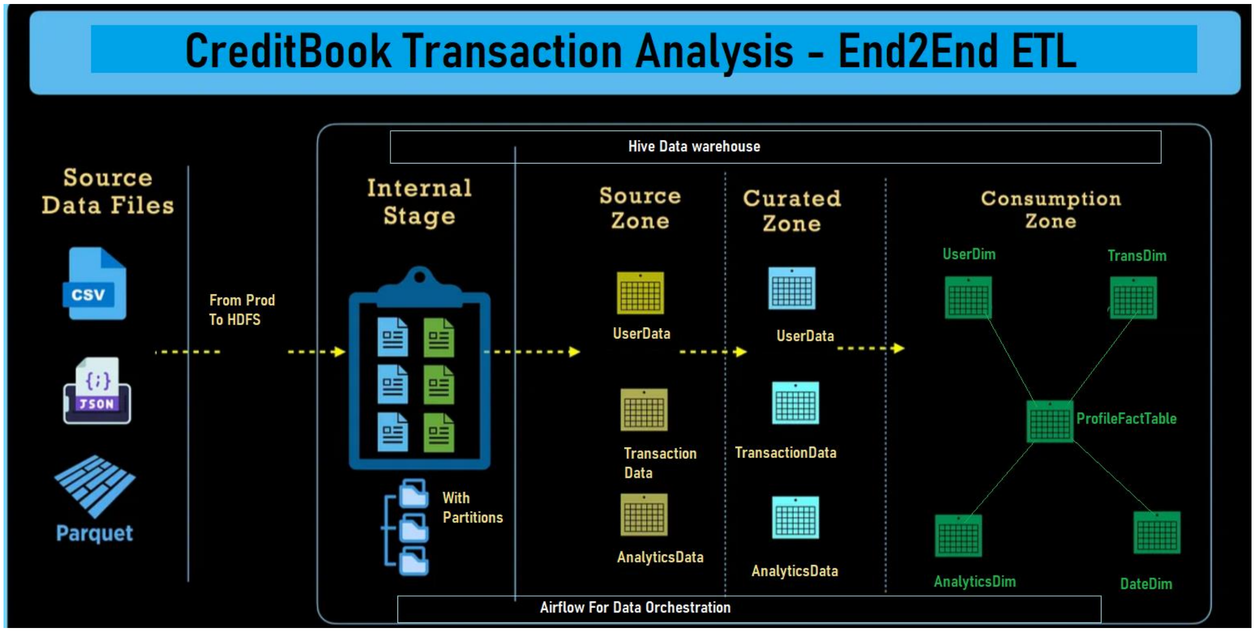
1. Introduction

This document outlines the design and implementation of an end-to-end ETL pipeline enriched with Apache Airflow integration. The pipeline is architected using Python, SQL, and various Apache tools, catering to the data processing needs of our client the CreditBook, with a focus on scalability, efficiency, and maintainability.

2. Infrastructure Setup

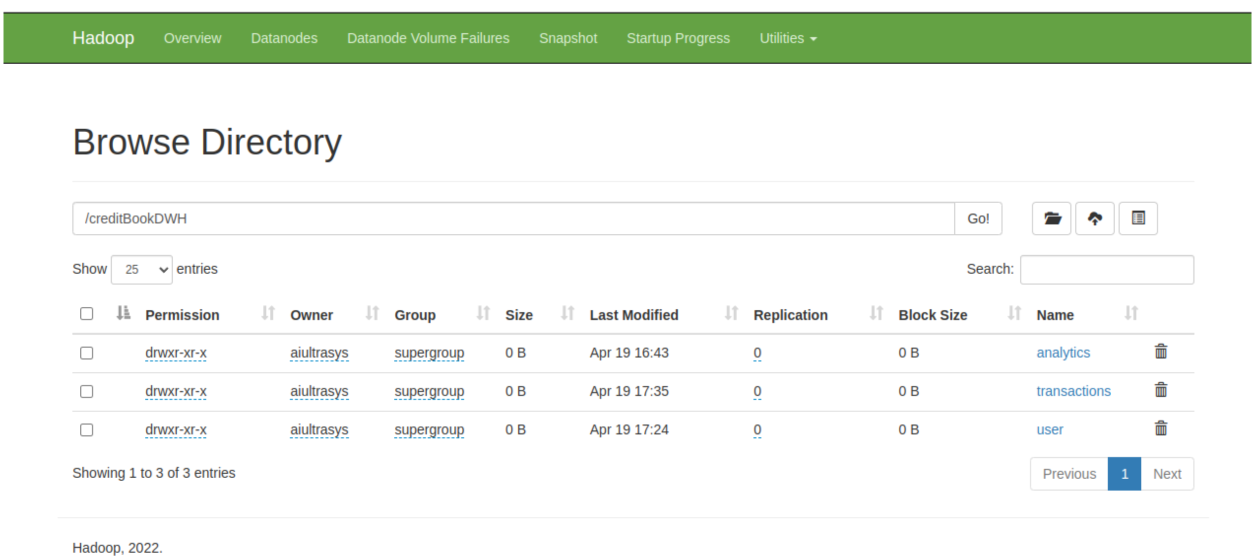
The infrastructure comprises several key components:

- **Storage Layer**: Hadoop Distributed File System (HDFS) is utilized for efficient storage and retrieval of large volumes of data.
- **Data Warehousing**: Apache Hive is employed as the data warehousing solution, providing structured querying capabilities.
- **ETL Engine**: Apache Spark serves as the ETL engine, enabling high-speed processing of data transformations.
- **Data Orchestration**: Apache Airflow is utilized for orchestrating the data pipeline, ensuring seamless execution of tasks.
- **Full Load & Incremental Load**: Using apache spark, I created a separate mechanism for full load and incremental load.
- **Query Monitoring**: Presto is integrated for monitoring query complexity and optimizing performance.



3. Data Ingestion

Data is ingested from the client's provided URLs and stored in HDFS in Parquet format with Snappy compression. This approach ensures efficient storage and retrieval of data, with partitions created based on date and time for enhanced query performance. Apache Spark is leveraged for data ingestion, facilitating parallel processing and scalability.



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 16:43	0	0 B	analytics
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:35	0	0 B	transactions
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:24	0	0 B	user

Browse Directory

/creditBookDWH/analytics | Go!

Show 25 entries   Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
	-rw-r--r--	aiultrasys	supergroup	0 B	Apr 19 16:43	3	128 MB	_SUCCESS	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	cryear=2020	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 16:43	0	0 B	cryear=2021	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 16:43	0	0 B	cryear=2022	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:41	0	0 B	cryear=2023	

Showing 1 to 5 of 5 entries   Previous 1 Next

Hadoop, 2022.

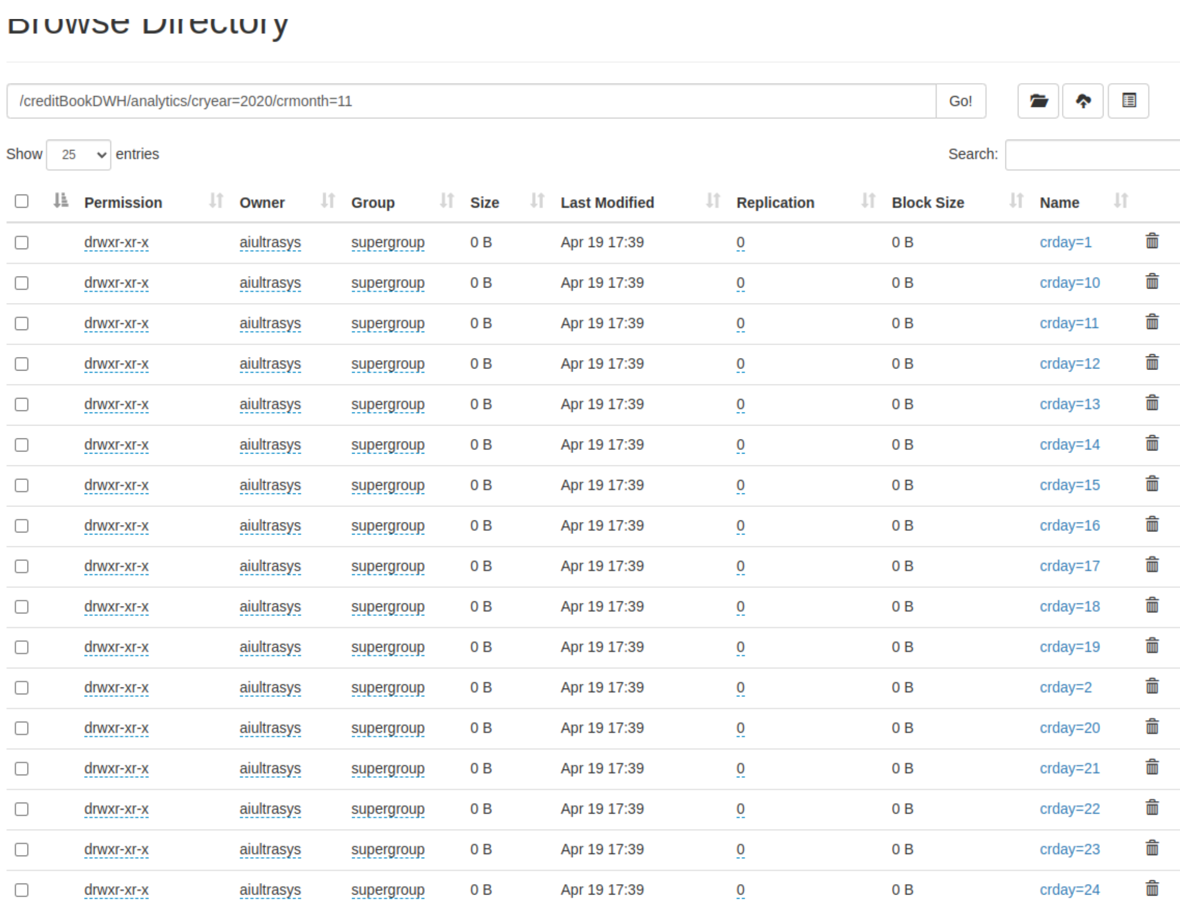
Browse Directory

/creditBookDWH/analytics/cryear=2020 | Go!

Show 25 entries   Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:38	0	0 B	crmonth=1	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crmonth=10	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crmonth=11	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crmonth=12	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:38	0	0 B	crmonth=2	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:38	0	0 B	crmonth=3	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:38	0	0 B	crmonth=4	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:38	0	0 B	crmonth=5	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:38	0	0 B	crmonth=6	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crmonth=7	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crmonth=8	
	drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crmonth=9	

Showing 1 to 12 of 12 entries   Previous 1 Next



Browse Directory

/creditBookDWH/analytics/cryear=2020/crmonth=11 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=1
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=10
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=11
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=12
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=13
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=14
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=15
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=16
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=17
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=18
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=19
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=2
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=20
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=21
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=22
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=23
drwxr-xr-x	aiultrasys	supergroup	0 B	Apr 19 17:39	0	0 B	crday=24

4. Layered Approach

The ETL pipeline is structured into three layers:

Source Layer

In the Source Layer, data is loaded into an external Hive table from HDFS. This enables seamless querying of raw data using Hive or Presto CLI, providing flexibility in data exploration.

Curated Layer

The Curated Layer focuses on data transformation and preparation. Various transformations such as cleaning, datatype correction, formatting, column creation, aggregation, and joins are performed to prepare the data for analysis. The transformed data is stored in an internal Hive table, ensuring curated data is readily accessible for downstream processing.

Consumption Layer

The Consumption Layer is where the transformed data is modeled for analytics purposes. A star schema is implemented, comprising fact and dimensional tables. Surrogate keys and foreign keys are established to maintain relationships between dimension and fact tables. Aggregate data is incorporated into the fact table to facilitate efficient analytics. Separate internal Hive tables are created for each dimension and fact table, organized based on primary keys, surrogate keys, and foreign keys.

```
CREATE EXTERNAL TABLE IF NOT EXISTS creditbook_sourcelayer.user_details (
    user_id STRING,
    business_id STRING,
    rating STRING,
    created_at STRING,
    processed_at timestamp
    )
partitioned by (cryear INT,crmonth INT,crday INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS PARQUET
LOCATION '/creditBookDWH/user'
;

create table IF NOT EXISTS creditbook_curatedlayer.user_details (
    user_id STRING,
    business_id STRING,
    rating STRING,
    created_at STRING,
    processed_at timestamp,
    signup_since_days INT,
    cryear INT,
    crmonth INT,
    crday INT);
```

```sql
create table IF NOT EXISTS creditbook_consumplayer.user_dim(
    user_id_pk INT PRIMARY KEY DISABLE NOVALIDATE,
    user_id STRING,
    business_id STRING,
    rating STRING,
    created_at STRING,
    processed_at STRING,
    signup_since_days STRING,
    cryear INT,
    crmonth INT,
    crday INT,
    isActive CHAR(1)
);
```

```
Query 20240422_144046_00149_w3iyx, RUNNING, 1 node
Splits: 18 total, 0 done (0.00%)
[Latency: client-side: 0:06, server-side: 63ms] [0 rows, 0B] [0 rows/s, 0B/s]

Query aborted by user
presto:creditbook_consumplayer> show schemas;
         Schema
-------------------------
 creditbook_consumplayer
 creditbook_curatedlayer
 creditbook_public
 creditbook_sourcelayer
```

```
Query 20240422_144341_00150_w3iyx, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
[Latency: client-side: 81ms, server-side: 75ms] [12 rows, 262B] [160 rows/s, 3.41KB/s]

presto:creditbook_consumplayer> use creditbook_sourcelayer;
USE
presto:creditbook_sourcelayer> show tables;
       Table
-------------------
 analytics_details
 trans_details
 user_details
(3 rows)

Query 20240422_144514_00154_w3iyx, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
[Latency: client-side: 87ms, server-side: 83ms] [3 rows, 138B] [36 rows/s, 1.62KB/s]

presto:creditbook_sourcelayer> 
```

```
presto:creditbook_sourcelayer> select count(*) from analytics_details;
 _col0
-------
 92869
(1 row)

Query 20240422_144637_00157_w3iyx, FINISHED, 1 node
Splits: 819 total, 819 done (100.00%)
[Latency: client-side: 0:01, server-side: 0:01] [92.9K rows, 4.36MB] [152K rows/s, 7.15MB/s]

presto:creditbook_sourcelayer> select count(*) from trans_details;
 _col0
--------
 696618
(1 row)

Query 20240422_144646_00158_w3iyx, FINISHED, 1 node
Splits: 950 total, 950 done (100.00%)
[Latency: client-side: 0:01, server-side: 0:01] [697K rows, 12.2MB] [837K rows/s, 14.7MB/s]

presto:creditbook_sourcelayer> select count(*) from user_details;
 _col0
-------
    64
(1 row)

Query 20240422_144657_00159_w3iyx, FINISHED, 1 node
Splits: 42 total, 42 done (100.00%)
[Latency: client-side: 373ms, server-side: 369ms] [64 rows, 52KB] [173 rows/s, 141KB/s]
```

```
presto:creditbook_curatedlayer> show tables;
       Table
-------------------
 analytics_details
 trans_details
 user_details
(3 rows)

Query 20240422_144748_00164_w3iyx, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
[Latency: client-side: 137ms, server-side: 133ms] [3 rows, 141B] [22 rows/s, 1.04KB/s]

presto:creditbook_curatedlayer> select count(*) from analytics_details;
 _col0
-------
 92869
(1 row)

Query 20240422_144759_00165_w3iyx, FINISHED, 1 node
Splits: 43 total, 43 done (100.00%)
[Latency: client-side: 103ms, server-side: 97ms] [92.9K rows, 13.8MB] [957K rows/s, 142MB/s]

presto:creditbook_curatedlayer> select count(*) from trans_details;
 _col0
--------
 696618
(1 row)

Query 20240422_144806_00166_w3iyx, FINISHED, 1 node
Splits: 28 total, 28 done (100.00%)
[Latency: client-side: 91ms, server-side: 87ms] [697K rows, 147MB] [8.01M rows/s, 1.65GB/s]

presto:creditbook_curatedlayer> select count(*) from user_details;
 _col0
-------
    25
(1 row)
```

```
presto:creditbook_consumplayer> show tables;
     Table
--------------
 analytics_dim
 date_dim
 profile_fact
 trans_dim
 user_dim
(5 rows)

Query 20240422_144943_00180_w3iyx, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
[Latency: client-side: 142ms, server-side: 136ms] [5 rows, 215B] [36 rows/s, 1.54KB/s]

presto:creditbook_consumplayer> select * from trans_dim limit 2;
 trans_id_pk |            transaction_id            |               user_id                |             business_id
-------------+--------------------------------------+--------------------------------------+-------------------------------------
      145549 | a22655b0-b571-4323-a172-be03a7f4894a | 49a9d296-e5a6-4c23-bf5e-9897dee07917 | 6fa4ccfc-2a88-4dc2-9692-400ef50623f
      145550 | 89ba45fe-c42b-49bd-abb5-2840e2af3cfc | 49a9d296-e5a6-4c23-bf5e-9897dee07917 | 6fa4ccfc-2a88-4dc2-9692-400ef50623f
(2 rows)

Query 20240422_145001_00181_w3iyx, FINISHED, 1 node
Splits: 22 total, 21 done (95.45%)
[Latency: client-side: 107ms, server-side: 103ms] [15.1K rows, 3.34MB] [147K rows/s, 32.4MB/s]
```

5. Data Modeling

The data modeling approach revolves around a star schema, which provides a structured and optimized data model for analytics. This schema consists of:

- **Fact Table**: Represents the core metrics of interest (e.g., transactions), containing surrogate keys and foreign keys to related dimension tables.
- **Dimension Tables**: Represent descriptive attributes (e.g., user details, date information), providing context to the facts. Each dimension table contains surrogate keys for efficient querying and joins.

```
Fact and Dimension Table:


1. Datedim
2. UserDim
3. TransDim
4. AnalyticsDim
5. ProfileFact


1. Datedim:

create table IF NOT EXISTS creditbook_consumplayer.date_dim(
date_id_pk INT PRIMARY KEY DISABLE NOVALIDATE,
trans_date                 date,
trans_year               INT,
trans_month              INT,
trans_quarter            INT,
trans_day                INT,
trans_dayofweek          INT,
trans_dayname            STRING,
trans_dayofmonth         INT,
trans_weekday            STRING);
```

2. UserDim

```
create table IF NOT EXISTS creditbook_consumplayer.user_dim(
    user_id_pk INT PRIMARY KEY DISABLE NOVALIDATE,
    user_id STRING,
    business_id STRING,
    rating STRING,
    created_at STRING,
    processed_at STRING,
    signup_since_days STRING,
    cryear INT,
    crmonth INT,
    crday INT,
    isActive CHAR(1)
);
```

3. TransDim

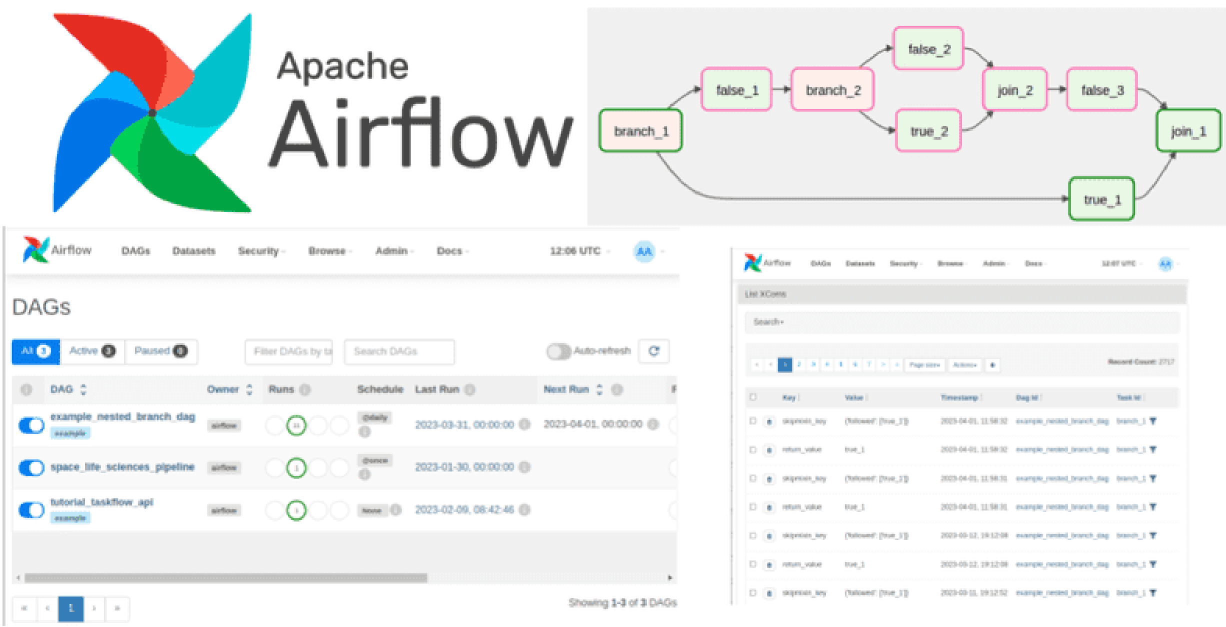
```
create table IF NOT EXISTS creditbook_consumplayer.trans_dim(

    trans_id_pk INT PRIMARY KEY DISABLE NOVALIDATE,
    transaction_id STRING,
    user_id STRING,
    business_id STRING,
    amount STRING,
    transaction_type STRING,
    type STRING,
    created_at STRING,
    processed_at timestamp,
    cryear INT,
    crmonth INT,
    crday INT,
    trans_date string,
    debit_count BIGINT,
    credit_count BIGINT,
    total_trans_count BIGINT,
    debitamount_inmillion DOUBLE,
    creditamount_inmillion DOUBLE,
    totaltransamount_inmillion DOUBLE,
    isActive CHAR(1)


);
```

```sql
create table IF NOT EXISTS creditbook_consumplayer.analytics_dim(

    analytics_id_pk INT PRIMARY KEY DISABLE NOVALIDATE,
    user_id STRING,
    event_date STRING,
    category STRING,
    mobile_brand_name STRING,
    mobile_model_name STRING,
    mobile_os_hardware_model STRING,
    operating_system STRING,
    operating_system_version STRING,
    city STRING,
    country STRING,
    appversion STRING,
    processed_at timestamp,
    cryear INT,
    crmonth INT,
    crday INT,
    isActive CHAR(1)
);
```

```sql
create table IF NOT EXISTS creditbook_consumplayer.profile_fact(
    profile_id_pk INT PRIMARY KEY DISABLE NOVALIDATE,
    user_id_fk INT,
    trans_id_fk INT,
    analytics_id_fk INT,
    date_id_fk INT,
    gmv_per_month DOUBLE,
    avgtranspermonth DOUBLE
);

alter table creditbook_consumplayer.profile_fact add constraint fk_user_id FOREIGN KEY (user_id_fk) REFERENCI

alter table creditbook_consumplayer.profile_fact add constraint fk_trans_id FOREIGN KEY (trans_id_fk) REFEREI

alter table creditbook_consumplayer.profile_fact add constraint fk_analytics_id FOREIGN KEY (analytics_id_fk

alter table creditbook_consumplayer.profile_fact add constraint fk_date_id FOREIGN KEY (date_id_fk) REFERENCI
```

6. Apache Airflow Integration

Apache Airflow is seamlessly integrated into the pipeline to orchestrate tasks and automate workflow execution. Directed Acyclic Graphs (DAGs) are created to define the workflow, ensuring tasks are executed sequentially or in parallel based on dependencies and schedules. This integration enhances pipeline reliability, scalability, and monitoring capabilities.



7. Full Load and Incremental Load Mechanisms

Within the Spark code, a separate mechanism is implemented to handle both full load and incremental load scenarios. This mechanism allows for flexibility in data loading strategies based on the requirements:

- **Full Load**: When the ETL code is executed with the "FullLoad" parameter, the entire dataset is processed and loaded from one layer to another. This ensures that all data is transferred and refreshed, useful for periodic updates or initial data setup.
- **Incremental Load**: In contrast, when the ETL code is executed with the "IncrementalLoad" parameter, only new or updated data is processed and loaded. This mechanism identifies changes since the last load and selectively transfers the relevant data, reducing processing time and resource utilization.

By implementing both full load and incremental load mechanisms, the ETL pipeline optimizes data processing efficiency and ensures that the most up-to-date information is available for analysis without unnecessary reprocessing of unchanged data.

8. Conclusion

In conclusion, the implemented ETL pipeline fulfills the client's requirements for transforming and aggregating their production database into a usable format for the data team. The modular design, coupled with the use of scalable technologies, ensures the pipeline's efficiency, flexibility, and maintainability.

```
presto:creditbook_consumplayer> select ccud.user_id, ccud.rating, ccud.signup_since_days,ccpf.gmv_per_month , ccpf.avgtranspermonth,
                             -> npdd.trans_date,npdd.trans_year,npdd.trans_month,npdd.trans_quarter,npdd.trans_weekday,npdd.trans_dayname,npdd.trans_dayofmonth,
                             -> cctd.amount, cctd.debit_count, cctd.credit_count, cctd.total_trans_count, cctd.debitamount_inmillion , cctd.creditamount_inmillion, cctd.totaltransamount_inmillion,
                             -> ccad.category, ccad.mobile_brand_name, ccad.city, ccad.appversion
                             -> from creditbook_consumplayer.profile_fact ccpf
                             -> inner join creditbook_consumplayer.user_dim ccud on ccpf.user_id_fk = ccud.user_id_pk
                             -> inner join creditbook_consumplayer.trans_dim cctd on ccpf.trans_id_fk = cctd.trans_id_pk
                             -> inner join creditbook_consumplayer.date_dim npdd on ccpf.date_id_fk = npdd.date_id_pk
                             -> inner join creditbook_consumplayer.analytics_dim ccad on ccpf.analytics_id_fk = ccad.analytics_id_pk
                             -> limit 20;
              user_id                | rating | signup_since_days |    gmv_per_month     |  avgtranspermonth  | trans_date | trans_year | trans_month | trans_quarter | trans_weekday | trans_dayname
-------------------------------------+--------+-------------------+----------------------+--------------------+------------+------------+-------------+---------------+---------------+--------------
 60b0bca5-aba2-4170-beb0-554b9e0b2a4e | 4      | 928               |          6.57996998E7 |  2154.188895072843 | 2022-08-14 |       2022 |           8 |             3 | Weekend       | Sunday
 65feb711-a05b-4eca-bfb5-ed7af23a2309 | 4      | 1333              |         1.059017875E8 | 3965.4679660001498 | 2021-11-16 |       2021 |          11 |             4 | Weekday       | Tuesday
 031e8f64-0e97-4f42-ab91-45569d114f0d | 5      | 1030              |        1.7495383505E8 |  5849.342529254431 | 2022-07-26 |       2022 |           7 |             3 | Weekday       | Tuesday
 cc0e78f9-305e-48e0-a4cf-72a79fb8a7d0 | 4      | 55                |         8.777229381E7 | 2599.5822121194174 | 2022-03-31 |       2022 |           3 |             1 | Weekday       | Thursday
 3febcb3d-3acc-4b14-95ef-caba16777a9e | 5      | 1101              | 8.904266475999999E7 | 2458.7233124395966 | 2022-04-28 |       2022 |           4 |             2 | Weekday       | Thursday
 4ddd310d-ccfa-4126-9c2a-1d37b07a36ed | 4      | 1216              |         9.391771534E7 |   2724.22669586657 | 2022-12-31 |       2022 |          12 |             4 | Weekend       | Saturday
 671d933f-f411-49f5-9a90-7875db66e984 | 5      | 1084              |          6.57996998E7 |  2154.188895072843 | 2022-08-18 |       2022 |           8 |             3 | Weekday       | Thursday
 65feb711-a05b-4eca-bfb5-ed7af23a2309 | 4      | 1333              |        1.4809863511E8 |  4154.355945748829 | 2022-05-14 |       2022 |           5 |             2 | Weekend       | Saturday
 45afcadc-3da4-44b6-96f0-b9a864d3b8f6 | 4      | 791               |         1.136326355E8 |    3241.8302949903 | 2022-09-16 |       2022 |           9 |             3 | Weekday       | Friday
 60b0bca5-aba2-4170-beb0-554b9e0b2a4e | 4      | 928               |         7.839444281E7 | 2085.3468148325487 | 2022-06-20 |       2022 |           6 |             2 | Weekday       | Monday
 3febcb3d-3acc-4b14-95ef-caba16777a9e | 5      | 1101              |         7.839444281E7 | 2085.3468148325487 | 2022-06-24 |       2022 |           6 |             2 | Weekday       | Friday
 dee57139-26ca-4dde-b9b1-ec42014cd394 | 4      | 1183              |         8.777229381E7 | 2599.5822121194174 | 2022-03-14 |       2022 |           3 |             1 | Weekday       | Monday
 b2ea88b8-a17b-4d89-913d-5a2fe9b4a277 | 4      | 1176              |        1.4809863511E8 |  4154.355945748829 | 2022-05-01 |       2022 |           5 |             2 | Weekend       | Sunday
 a50fc2e9-1454-4754-96d9-86a0e0fdc36c | 4      | 1155              |          6.6117154E7 | 2216.9853468799247 | 2022-02-03 |       2022 |           2 |             1 | Weekday       | Thursday
 a8a45c83-ae63-40ef-8fc0-4b958fdc90bc | 4      | 1148              |         9.391771534E7 |   2724.22669586657 | 2022-12-01 |       2022 |          12 |             4 | Weekday       | Thursday
 hrV88QmZteSZwb7ZH1pQo2lJT8i2         | 4      | 537               |        1.1045658175E8 |  2060.834018993246 | 2022-11-02 |       2022 |          11 |             4 | Weekday       | Wednesday
 6ab12451-393e-401a-a111-e7f2a47178f8 | 4      | 1231              | 8.904266475999999E7 | 2458.7233124395966 | 2022-04-22 |       2022 |           4 |             2 | Weekday       | Friday
 671d933f-f411-49f5-9a90-7875db66e984 | 5      | 1084              | 8.904266475999999E7 | 2458.7233124395966 | 2022-04-29 |       2022 |           4 |             2 | Weekday       | Friday
 7e848c3c-0f00-4c6a-98e7-65734cadf6c4 | 5      | 914               |        5.9137735913E8 | 18709.736747975196 | 2022-01-23 |       2022 |           1 |             1 | Weekend       | Sunday
 49a9d296-e5a6-4c23-bf5e-9897dee07917 | 5      | 925               |         9.191531319E7 | 2755.9160826936914 | 2023-01-13 |       2023 |           1 |             1 | Weekday       | Friday
(20 rows)
(END)
```