

**Capstone Project:**

**Supplier Performance Analysis Design/Approach Document**

**Git Repo Location:** https://github.com/ShaktiGoyal123/data-eng-capstone-project-teameinstein.git

**Slack Channel:** https://join.slack.com/share/enQtNzUwNDY2OTkyMTA5NC02NTdjYjE5ODBlZTg4MGVlZjNjMWFlN2NmYmExYmEwMmRiODZiN2JjMGU4NDMwYjgyN2VmNTIwMTBmNzIwNDJj

**Database Connection Details:**

**S3 Location Details:**

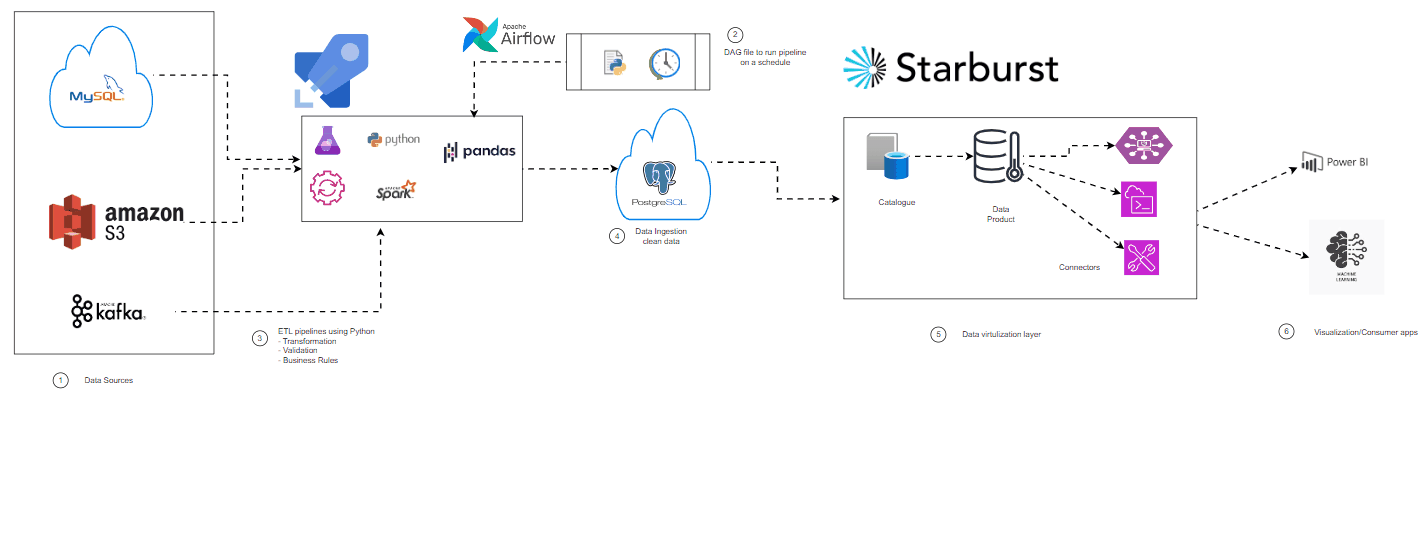
**Mongodb Location Details:**

**Datasets fetched from 3rd party sources like Kaggle**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Responsible for** | **Role** | **Comment** |
| **Goyal, Shakti** |  | **Lead** |  |
| **P, Saravanakumar** |  | **Data Engineer** |  |
| **Sarashetti, Praful** |  | **Data Engineer** |  |
| **Subramanian, Ramesh** |  | **Data Engineer** |  |
| **Revanth Nandhanur** |  | **Data Engineer** |  |
| **Gokavarapu, Parvathi** |  | **Data Engineer** |  |

**Project Components**

**Use case Diagram**



1. **Data Ingestion**:  
   **Sources**: CSV, SQL Tables, Iceberg Tables  
   **Technologies**: S3 (MinIO), Spark, Python, Postgres

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Source** | **Technology** |
| Supplier Information Data | SQL Table | Postgres |
| Purchase Order Data | SQL Tables | Postgres |
| Delivery Data | S3 (MinIO) | CSV |
| Quality Control Data | S3 (MinIO) | CSV |
| Lead Time Data | S3 (MinIO) | CSV |
| Return Data | S3 (MinIO) | CSV |
| Inventory Management Data | S3 (MinIO) | CSV |
| Cost Data | S3 (MinIO) | CSV |
| Invoice Data | SQL Tables | MYSQL |
| Compliance Data | SQL Table | Postgres |
| Performance Feedback Data | SQL Table | Postgres |

**2.Data Processing**:

|  |  |  |
| --- | --- | --- |
| **Pipeline** | **Type** | **Technology** |
| Python | Batch | Pandas & Spark |

**3. Data Storage (Sink)**:

|  |  |  |
| --- | --- | --- |
| **Pipeline** | **Destination** | **Format** |
| Supplier Information Data | Starburst | catalog |
| Purchase Order Data | Postgres | Table |
| Delivery Data | Postgres | Catalog Iceberg |
| Quality Control Data | Postgres | Catalog Iceberg |
| Lead Time Data | Postgres | Catalog Iceberg |
| Return Data | Postgres | Catalog Iceberg |
| Inventory Management Data | Postgres | Catalog Iceberg |
| Cost Data | Postgres | Catalog Iceberg |
| Invoice Data | Postgres | Table |
| Compliance Data | Starburst | Catalog |
| Performance Feedback Data | Starburst | Catalog |

**4.Data Analysis**:

|  |  |  |
| --- | --- | --- |
| **Dataset** | ***Business Logic/Transformation*** | **Data Model** |
| Supplier Information Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Purchase Order Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Delivery Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Quality Control Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Lead Time Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Return Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Inventory Management Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Cost Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Invoice Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |
| Compliance Data | Remove all the null records.  No Special Characters  Date Format validation  Non-Zero values for prices |  |

**5.Data Visualization**:

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Report/Chart Name** | **Report Type** |
| Supplier Performance Data Mart | Supplier Performance Analysis | Bar Chart |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

6. Packaging/Deployment (Workflows)

|  |  |  |
| --- | --- | --- |
| **Workflow Name** | **Tasks** | **Dependency** |
| Ingesting the data using Pipeline | Data Ingestion |  |
| Data Refresh from MINIO to PostgreSQL | Data Refresh | Data Ingestion |

**Issue Tracker**

<<document the issues which are faced during the capstone project development>>

|  |  |  |
| --- | --- | --- |
| **Description** | **Type Of issue faced** | **Solution** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |