**On-Platform Data Pipeline-**

**Content**: -

|  |  |
| --- | --- |
| **Serial No** | **Content Available** |
| 1. | Introduction |
| 2. | List Of On-Platform Data Sources Available |
| 3. | What is Data Ingestion Process |
| 4. | What is Data Processing Pipeline |
| 5. | Dataset with their Respective Table Description |

**Introduction: -**

An On-Platform data source providing social media data refers to an interface or system that enables the retrieval, collection, and analysis of data generated from social media platforms. This can include APIs, data streams, and third-party tools.

These data sources serve as critical components for building applications, conducting research, and deriving insights from social media interactions and trends.

Data is a two-part structure that ensures a clear separation of concerns between data ingestion and processing. Each part can be scaled and optimized independently, allowing for a robust data pipeline that can handle diverse data sources and processing needs. The two different Parts are DIP (Data Ingestion Process) and DPP (Data Processing Process.)

**List Of On-Platform Data Sources Available: -**

|  |  |
| --- | --- |
| **Serial No** | **On-Platform sources** |
| 1. | CMP/Cerebro |
| 2. | CAP |
| 3. | GAM |
| 4. | salesforce |
| 5. | Workfront |
| 6. | Criteo |
| 7. | Citrus |
| 8. | Campaign Aggregate Table |
| 9. | Rubix |
| 10. | Store Presentation/Signage |
| 11. | Instore/ServiceHub |

**What is Data Ingestion Process**: -

A data ingestion pipeline typically focuses on efficiently collecting, processing, and managing data, especially in the context of a cloud environment and Elasticsearch for search and analytics. Below is a structured outline of such a pipeline:

### **Data Ingestion Process Overview**

1. **Data Sources**: Identify the various data sources from which data will be ingested.
2. **Data Ingestion Tools**: Use tools to collect data and push it into a processing layer.
3. **Data Processing**: Transform and prepare the ingested data for storage and analysis.
4. **Data Storage**: Store the processed data in a suitable format, often in Elasticsearch.
5. **Monitoring and Management**: Utilize tools like Cerebro for managing Elasticsearch and monitoring the pipeline.

**What is a Data Processing Pipeline-**

Creating a data processing pipeline involves integrating various components to handle data ingestion, processing, storage, and analysis. Below is a high-level overview of how such a pipeline can be structured, assuming you’re referring to a typical setup for data processing.

### **Data Processing Pipeline Overview**

1. **Data Ingestion**: Collect data from various sources.
2. **Data Processing**: Transform and analyze the ingested data.
3. **Data Storage**: Store processed data in a suitable format.
4. **Data Visualization/Analysis**: Analyze or visualize the data for insights.

**Data Set with their Respective Tables Description -**

**1.CMP/Cerebro-**

**List Of tables Under Cerebro: -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Schema Name** | **Table Name** | **SLA** | **Load\_Time** | **Load\_Frequency** | **frequecy\_of\_ingestion\_with\_time** |
| prd\_mdf\_fnd | cerebro\_campaign\_adgroups | 10:00 PM | 06:30 AM CST | Daily | twice a day 5:00 and 17:00 |
| cerebro\_campaign\_details | 8:00 PM | 05:00 AM | Daily | twice a day 3:30 and 15:30 |
|  | salesforce\_campaign\_details | 10:00 PM | 4:45 AM | Daily | twice a day 4:00 and 16:00 |
|  | Campaign\_Media\_Brief | 10:00 PM | 06:15 AM | Daily | twice a day 4:30 and 16:30 |
|  | campaign\_service\_providers | 10:00 PM | 05:15 AM | Daily | twice a day 3:30 and 15:30 |
|  | Campaign\_rate\_card | 10:00 PM | 5:00 AM | Daily | twice a day 5:00 and 17:00 |
|  | Tactics\_template | 10:00 PM | 4:00 AM | Daily | twice a day 6:00 and 16:00 |
|  | PBR\_deal\_updates | |  | | --- | | 12:00 PM | | 8:00 PM | Daily | twice a day 4:00 and 16:00 |
|  | CAP\_audiences\_updates | 9:00 AM | 02:15 PM | Daily | twice a day 10:30 and 22:30 |

**Data Ingestion Process- CMP/ Cerebro- Kafka->Redex->HDFS**

The Ingestion process remains the same for all the tables in the above-mentioned dataset. Creating a data ingestion pipeline that moves data from Kafka to Redex and then to HDFS (Hadoop Distributed File System) involves several steps and components. Below is a high-level overview of how this pipeline is set up.

### **Data Ingestion Process Overview**

1. **Kafka**: Acts as the initial data source, collecting and streaming data in real time.
2. **Redex**: Temporarily stores data for fast access and processing.
3. **HDFS**: Serves as the final storage layer for batch processing and analysis.

This pipeline provides a scalable way to ingest data in real time from Kafka, store it temporarily in Redis for quick access, and then batch it to HDFS for long-term storage and analysis. You can expand this pipeline further based on specific use cases, such as adding error handling, data transformation, or additional storage layers.

**Data Processing Pipeline- CMP/ Cerebro-**

The processing Process Differs for all the tables mentioned above-

1. **Cerebro Campaign Ad Groups-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svrdehdpd | Ami Lathia | 412 |
| svrikbrd | L Shekhar | 197 |
| svsophdp | Srivatchala Thilagaraj | 175 |
| svdomhdp | Vijaya Kumari | 158 |
| svmdlhdp | Vijaya Kumari | 133 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th, 2024-January 20th, 2025-none

\*Average Start Time In CST- **18:59:52**

\*Average END Time In CST- **19:18:46**

**\*Average Run Time In CST- 19 Min**

\*Average Start Time In UTC- **00:59:52**

\*Average END Time In UTC- **01:18:46**

**\*Average Run Time In UTC-19 min**

\*Average Throughput-

Below is a breakdown of the column description: -

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| placement\_id | Unique identifier for each ad placement. |
| ad\_group\_id | Identifier for the ad group associated with the placement. |
| placement\_name | Name of the ad placement. |
| publisher\_name | Name of the publisher where the ad is displayed. |
| campaign\_id | Identifier for the advertising campaign. |
| tactic\_template\_id | ID representing the marketing tactic used. |
| impression\_buffer\_percent | Buffer percentage applied to impressions for pacing. |
| adjusted\_impressions | Adjusted number of ad impressions. |
| impressions\_q | Total number of impressions. |
| impressions\_with\_buffer | Impressions adjusted with the buffer applied. |
| cost\_a | Cost attributed to the ad placement. |
| publisher\_media\_cost\_a | Cost incurred by the publisher for media. |
| advertiser\_cost\_a | Cost charged to the advertiser. |
| budget\_a | Allocated budget for the placement. |
| placement\_start\_date | Start date of the ad placement. |
| placement\_end\_date | End date of the ad placement. |
| placement\_settings\_platform\_name | The platform where the ad is placed (e.g., Mobile, Desktop). |
| placement\_settings\_rate\_type | The type of rate applied (e.g., CPM, CPD). |
| inventory\_size\_targeting\_arr | The targeted ad inventory sizes. |
| device\_targeting\_arr | Targeted devices for the ad placement. |

1. **Cerebro Campaign Details-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svrdehdpd | Ami Lathia | 412 |
| svrikbrd | L Shekhar | 197 |
| svsophdp | Srivatchala Thilagaraj | 175 |
| svdomhdp | Vijaya Kumari | 158 |
| svmdlhdp | Vijaya Kumari | 133 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

\*Average Start Time In CST- **13:08:52**

\*Average END Time In CST- **13:19:08**

**\*Average Run Time In CST- 10 Min**

\*Average Start Time In UTC- **19:08:52**

\*Average END Time in UTC **19:08:52**

**\*Average Run Time In UTC-10 min**

\*Average Throughput-

Below is a breakdown of the column descriptions:-

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| package\_id | Unique identifier for the campaign package |
| campaign\_id | Unique identifier for the campaign |
| campaign\_n | Name of the campaign |
| source\_n | Source or channel of the campaign |
| campaign\_type | Classification of the campaign (Standalone, etc.) |
| salesforce\_campaign\_n | Campaign name as recorded in Salesforce |
| salesforce\_campaign\_id | Unique Salesforce campaign ID |
| work\_front\_project\_id | Identifier for the Workfront project associated with the campaign |
| media\_brief\_id | ID of the media brief |
| budget\_a | Allocated budget for the campaign |
| campaign\_objective | Goal of the campaign (e.g., Conversions, Awareness) |
| category\_n | Industry category of the campaign |
| advertiser\_n | Name of the advertiser |
| brand\_list\_arr | List of brands involved in the campaign |
| licensed\_brand\_list\_arr | Licensed brands associated with the campaign |
| account\_manager\_id | ID of the account manager |
| account\_manager\_n | Name of the account manager |
| account\_manager\_email | Email of the account manager |
| campaign\_start\_utc\_ts | Start date and time of the campaign (UTC) |
| campaign\_end\_utc\_ts | End date and time of the campaign (UTC) |
| platform\_list\_arr | Platforms where the campaign is executed |
| tcin\_arr | Target item numbers associated with the campaign |
| licensed\_tcin\_arr | Licensed Target item numbers for the campaign |
| campaign\_status | Status of the campaign (Completed, Active, etc.) |
| rate\_tier | Pricing tier applied to the campaign |
| original\_rate\_tier | Initially assigned pricing tier |
| tactic\_list\_arr | List of tactics used in the campaign |
| campaign\_fiscal\_year | Fiscal year in which the campaign runs |
| planning\_fiscal\_year | Fiscal year in which the campaign was planned |
| last\_update\_utc\_ts | Timestamp of the last update made to the campaign |
| created\_utc\_ts | Timestamp of when the campaign was created |
| last\_status\_update | Last recorded status update of the campaign |
| is\_optimization\_available\_f | Flag indicating if optimization is available |
| media\_analyst\_id | ID of the media analyst assigned |
| media\_analyst\_n | Name of the media analyst |
| media\_analyst\_email | Email of the media analyst |
| trading\_analyst\_n | Name of the trading analyst |
| trading\_analyst\_email | Email of the trading analyst |
| levels\_of\_service | Service level classification |
| optimization\_eligibility | Eligibility criteria for optimization |
| optimization\_auto\_approve\_exclusion\_reason | Reason for exclusion from auto-approval of optimization |
| is\_optimization\_auto\_approved\_f | Flag indicating if optimization is auto approved |
| is\_inventory\_crunch\_available\_f | Flag for inventory crunch availability |
| is\_licensed\_list\_required\_f | Flag indicating if a licensed list is needed |
| optimization\_settings\_arr | Optimization settings applied |
| load\_d | Load date of the campaign record |
| report\_d | Report date of the campaign record |

1. **Salesforce campaign details-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svpachds | Vijaya Kumari | 415 |
| svpachdp | Vijaya Kumari | 356 |
| svmdedmp | Hemendra Singh | 266 |
| svfdsroundelhdp | Narendra Yaddula | 14 |
| z00ct7x | Karen Lin | 9 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

\*Failure Rate- Dated October 20th, 2024-January 20th, 2025-none

\*Average Start Time In CST- **03:48:16**

\*Average END Time In CST- **04:02:31**

**\*Average Run Time In CST- 14 Min**

\*Average Start Time In UTC- **09:48:16**

\*Average END Time In UTC- **10:02:31**

**\*Average Run Time In UTC-14 min**

\*Average Throughput-

Below is a breakdown of the column descriptions.

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| campaign\_market\_place\_id | Unique identifier for the campaign in the marketplace |
| salesforce\_id | Unique Salesforce ID for tracking the campaign |
| salesforce\_campaign\_name | Name of the campaign as recorded in Salesforce |
| budget\_a | Budget allocated for the campaign |
| campaign\_status | Status of the campaign (Active, Completed, etc.) |
| media\_brief\_id | Identifier for the media brief linked to the campaign |
| package\_id | Unique identifier for the package associated with the campaign |
| objective\_details | Primary goal of the campaign (e.g., Conversions) |
| category\_n | Category of the product being advertised (e.g., Beauty, Grocery) |
| account\_manager\_id | Unique identifier for the account manager handling the campaign |
| account\_manager\_name | Name of the account manager |
| account\_manager\_email | Email of the account manager |
| advertiser\_list\_arr | List of advertisers associated with the campaign |
| start\_lcl\_ts | Local start date and time of the campaign |
| end\_lcl\_ts | Local end date and time of the campaign |
| last\_update\_lcl\_ts | Timestamp of the last update made to the campaign |
| load\_d | Load date of the campaign record |
| report\_d | Date when the campaign data was reported |

1. **Campaign\_Media\_Brief-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| z009v0f | Tushara Nivarthi | 193 |
| svmdedmp | Hemendra Singh | 90 |
| z00fh9k | Nithish Loganathan | 7 |
| z009498 | Amal Abdussalam PT | 3 |
| z00bj0y | Rishabh Kaila | 3 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

\*Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **07:33:17**

\*Average END Time In CST- **07:41:08**

**\*Average Run Time In CST- 8 Min**

\*Average Start Time In UTC- **13:33:17**

\*Average END Time In UTC- **13:41:08**

**\*Average Run Time In UTC-8 min**

\*Average Throughput-

1. **Campaign\_Service\_Provider-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svopthdp | Narendra Yaddula | 647 |
| svopthds | Narendra Yaddula | 542 |
| svmdedmp | Hemendra Singh | 93 |
| z079730 | Sasidhar Myneni | 11 |
| z00fd1j | Sourav Raj | 10 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

\*Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **07:35:54**

\*Average END Time In CST- **08:02:01**

**\*Average Run Time In CST- 26 Min**

\*Average Start Time In UTC- **13:35:54**

\*Average END Time In UTC- **14:02:01**

**\*Average Run Time In UTC-26 min**

\*Average Throughput-

1. **Campaign\_Rate\_Card-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| svmdedmp | Hemendra Singh | 91 |
| z00fd1j | Sourav Raj | 7 |
| svmdedmd | Hemendra Singh | 2 |
| z00ffrs | Anup Jee | 2 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

\*Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **18:06:41**

\*Average END Time In CST- **18:25:41**

**\*Average Run Time In CST- 19 Min**

\*Average Start Time In UTC- **00:06:41**

\*Average END Time In UTC- **00:25:41**

**\*Average Run Time In UTC-19 min**

\*Average Throughput-

1. **Tactic\_Templates-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svopthdp | Narendra Yaddula | 659 |
| svopthds | Narendra Yaddula | 542 |
| svpachds | Vijaya Kumari | 400 |
| svpachdp | Vijaya Kumari | 368 |
| svrdehdpd | Ami Lathia | 301 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

\*Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **03:41:40**

\*Average END Time In CST- **04:15:12**

**\*Average Run Time In CST- 34 Min**

\*Average Start Time In UTC- **09:41:40**

\*Average END Time In UTC- **10:15:12**

**\*Average Run Time In UTC-34 min**

\*Average Throughput-

1. **PBR\_Deal\_Updates-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svrndlmsrmthdp | Hemendra Singh | 632 |
| svrndlmsrmthds | Hemendra Singh | 429 |
| svmdedmp | Hemendra Singh | 82 |
| svcycttdowp | Nomvelo Moyo | 76 |
| svdomhdp | Vijaya Kumari | 72 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **21:30:25**

\*Average END Time In CST- **21:40:55**

**\*Average Run Time In CST- 10 Min**

\*Average Start Time In UTC- **03:30:25**

\*Average END Time In UTC- **03:40:55**

**\*Average Run Time In UTC-10 min**

\*Average Throughput-

1. **CAP\_Audience\_Updates-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svpachds | Vijaya Kumari | 425 |
| svpachdp | Vijaya Kumari | 359 |
| svrpihds | L Shekhar | 323 |
| svaudience | Jason Lee | 159 |
| svrpihdp | L Shekhar | 124 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **02:46:04**

\*Average END Time In CST- **03:09:18**

**\*Average Run Time In CST- 23 Min**

\*Average Start Time In UTC- **08:46:04**

\*Average END Time In UTC- **09:09:18**

**\*Average Run Time In UTC-23 min**

\*Average Throughput-

**2.GAM (Google Ads Manager)-**

**List Of tables Under GAM: -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Schema Name** | **Table Name** | **SLA** | **Load\_Time** | **Load\_Frequency** | **frequecy\_of\_ingestion\_with\_time** |
| prd\_mdf\_fnd | GAM\_API\_Actuals | 8AM CST | |  | | --- | | 4AM CST | | Daily | 1, 8AM CST |
| GAM\_Availability\_forecast | 8PM CST | |  | | --- | | 8AM CST | | Daily | 1, 8AM CST |
|  | GAM\_Delivery\_forecast | 8PM CST | |  | | --- | | 8AM CST | | Daily | 1, 8AM CST |
|  | GAM\_Optimus\_Day\_Line\_item\_forecast\_agg |  | |  | | --- | | 8AM CST | | Daily | 1, 8AM CST |
|  | GAM\_Delivery | 7.30PM CST | |  | | --- | | 4.15AM CST | | Daily | thrice a day and 4.15AM CST, 2.15PM CST and 3.15PM |

**1.) GAM API Actuals- Data Ingestion Process- GAM- GAM api → common api ingestion wrapper → /user/SVMDEDMP/hive/gam\_api\_actuals/gam\_daily\_land/**

## **Data Ingestion Process Overview- GAM API Actuals**

Data ingestion pipelines are crucial for automating the collection, transformation, and storage of data from various sources into a centralized data platform. In this context, we're focusing on ingesting data from **Google Ad Manager (GAM) API** into a **Hive-based data lake** using a **Common API Ingestion Wrapper**.

This pipeline is designed to:

* **Automate Daily Data Pulls:** Extract GAM actuals data daily, ensuring timely insights.
* **Standardize Ingestion Process:** Utilize a reusable wrapper for consistent API integration.
* **Ensure Data Quality:** Validate data for completeness, availability, and accuracy.
* **Optimize Query Performance:** Store data in partitioned formats for faster analytics.

1. **GAM API:** Extracts daily actuals data .
2. **Common API Ingestion Wrapper:** Standardizes, validates, and automates data ingestion
3. **/user/SVMDEDMP/hive/gam\_api\_actuals/gam\_daily\_land/:** Stores raw data in a partitioned structure for efficient querying and processing in Hive.

This end-to-end ingestion pipeline automates the extraction, transformation, and storage of GAM actuals data using a **Common API Ingestion Wrapper**. It provides a scalable, efficient, and secure approach to data integration while ensuring data quality and performance optimization.

**Data Processing Pipeline- GAM API Actuals-**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svpachds | Vijaya Kumari | 408 |
| svpachdp | Vijaya Kumari | 401 |
| svmdedmp | Hemendra Singh | 81 |
| svmdedmd | Hemendra Singh | 23 |
| z00fd1j | Sourav Raj | 9 |
|  |  |  |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **19:31:22**

\*Average END Time In CST- **23:41:56**

**\*Average Run Time In CST- 251 Min**

\*Average Start Time In UTC- **01:31:22**

\*Average END Time In UTC- **05:41:56**

**\*Average Run Time In UTC-251 min**

\*Average Throughput-

**2.) GAM Availability Forecast → common api ingestion wrapper → /user/SVMDEDMP/hive/gam\_api\_forecasts/gam\_daily\_land/**

## **Data Ingestion Process Overview- GAM Availability Forecast**

Data ingestion pipelines are critical for automating the extraction, transformation, and storage of data from various sources into a centralized data platform. In this case, the pipeline is designed to ingest **Google Ad Manager (GAM) forecast data** into a **Hive-based data lake** using a **Common API Ingestion Wrapper**.

This pipeline enables:

* **Automated Daily Data Pulls:** Ensures timely and accurate forecast updates.
* **Standardized Ingestion Process:** Reusable wrapper for consistent API integration.
* **Data Quality Assurance:** Automated validation for completeness, availability, and accuracy.
* **Optimized Query Performance:** Partitioned storage for faster analytics and reporting.

1. **GAM API:** Extracts daily forecast data
2. **Common API Ingestion Wrapper:** Standardizes, validates, and automates data ingestion
3. **user/SVMDEDMP/hive/gam\_api\_forecasts/gam\_daily\_land/:** Stores raw data in a partitioned structure for efficient querying and processing in Hive.

**Data Processing Pipeline- GAM Availability Forecast**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| z00gbl3 | Shreya Roy | 5 |
| svmdedmd | Hemendra Singh | 3 |
| svopthdp | Narendra Yaddula | 2 |
| svrpihds | L Shekhar | 1 |
| z00c526 | Reenu Kurian | 1 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-noneI(Lat Updated on 11th Nov)

\*Average Start Time In CST- **15:25:48**

\*Average END Time In CST- **00:47:02**

**\*Average Run Time In CST- 561 Min**

\*Average Start Time In UTC- **20:25:48**

\*Average END Time In UTC- **05:47:02**

**\*Average Run Time In UTC-561 min**

\*Average Throughput-

**3.) GAM\_delivery\_forecasts- The Ingestion Pipeline remains the same as stated above.**

**Data Processing Pipeline- GAM Delivery Forecast**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| z00fd1j | Sourav Raj | 7 |
| svopthdp | Narendra Yaddula | 3 |
| svmdedmp | Hemendra Singh | 2 |
| z00dw51 | VEERESH ALLI | 2 |
| z079730 | Sasidhar Myneni | 2 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-noneI(Lat Updated on 11th Nov)

\*Average Start Time In CST- **15:25:48**

\*Average END Time In CST- **00:47:02**

**\*Average Run Time In CST- 561 Min**

\*Average Start Time In UTC- **20:25:48**

\*Average END Time In UTC- **05:47:02**

**\*Average Run Time In UTC-561 min**

\*Average Throughput-

**4.) GAM\_Optimus\_Day\_Line\_Forecast\_Agg- The Ingestion Pipeline remains the same as stated above.**

**Data Processing Pipeline- GAM\_Optimus\_Day\_Line\_Forecast\_Agg**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svrpihds | L Shekhar | 342 |
| svrpihdp | L Shekhar | 121 |
| svmdedmp | Hemendra Singh | 85 |
| svopthdp | Narendra Yaddula | 8 |
| z00c3l0 | Yogesh Soniwal | 7 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-noneI (Lat Updated on 11th Nov)

\*Average Start Time In CST- **15:25:48**

\*Average END Time In CST- **00:47:02**

**\*Average Run Time In CST- 561 Min**

\*Average Start Time In UTC- **20:25:48**

\*Average END Time In UTC- **05:47:02**

**\*Average Run Time In UTC-561 min**

\*Average Throughput-

**5.) GAM Delivery- GAM api → common api ingestion wrapper → /user/SVMDEDMP/hive/gam\_api\_delivery/gam\_api\_delivery\_daily\_land/**

## **Data Ingestion Process Overview- GAM Delivery**

This pipeline is designed to ingest **Google Ad Manager (GAM) delivery data** into a **Hive-based data lake** using a **Common API Ingestion Wrapper**. It provides:

* **Automated Daily Data Pulls:** Captures delivery metrics like impressions, clicks, and revenue.
* **Standardized Ingestion Process:** Ensures consistent data extraction, validation, and storage.
* **Data Quality Assurance:** Validates completeness, availability, and accuracy before storage.
* **Optimized Performance:** Partitioned storage for faster analytics and reporting.

1. **GAM API:** Extracts daily delivery data
2. **Common API Ingestion Wrapper:** Standardizes, validates, and automates data ingestion
3. **user/SVMDEDMP/hive/gam\_api\_delivery/gam\_api\_delivery\_daily\_land/:** Stores raw data in a partitioned structure for efficient querying and processing in Hive.

**Data Processing Pipeline- GAM Delivery**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svmdedmp | Hemendra Singh | 262 |
| z00c3l0 | Yogesh Soniwal | 17 |
| svmdedmd | Hemendra Singh | 9 |
| z079730 | Sasidhar Myneni | 4 |
| z00d0zg | Preetham Gittagandla | 2 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-noneI(Lat Updated on 11th Nov)

\*Average Start Time In CST- **06:09:34**

\*Average END Time In CST- **06:40:11**

**\*Average Run Time In CST- 31 Min**

\*Average Start Time In UTC- **12:09:34**

\*Average END Time In UTC- **12:40:11**

**\*Average Run Time In UTC-31 min**

\*Average Throughput-

**3.Salesforce-**

**List Of tables Under Salesforce: -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Schema Name** | **Table Name** | **SLA** | **Load\_Time** | **Load\_Frequency** | **frequecy\_of\_ingestion\_with\_time** |
| prd\_mdf\_fnd | |  | | --- | | salesforce\_campaign | | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | salesforce\_campaign\_product | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | |  | | --- | | salesforce\_opportunity | | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | salesforce\_revenue | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | |  | | --- | | salesforce\_brand | | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd  prd\_mdf\_fnd | |  | | --- | | salesforce\_brand\_opportunity | | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | salesforce\_brand\_agency | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | salesforce\_account | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | |  | | --- | | salesforce\_class | | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | |  | | --- | | salesforce\_contact | | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | salesforce\_vendor | 1:00 PM CST | |  | | --- | | 10:00 PM IST | | Daily | 1, 10PM IST |
| prd\_mdf\_fnd | salesforce\_user | NA | 10:00 PM IST | Daily | 1, 10PM IST |

**1.) Salesforce\_Campaign- Data Ingestion Process- salesforce→TAP->/user/SVMDEDMP/hive/salesforce/campaign/campaign\_report\_land.**

## **Data Ingestion Process Overview- Salesforce\_Campaign**

This pipeline is designed to ingest **Salesforce Campaign Reports** into a **Hive-based data lake** using **TAP (Transfer and Processing)**. It provides:

* **Automated Data Extraction:** Captures campaign metrics like impressions, clicks, and conversions.
* **Standardized Ingestion Process:** Ensures consistent extraction, transformation, and storage.
* **Data Quality Assurance:** Validates completeness, availability, and accuracy before storage.
* **Optimized Performance:** Partitioned storage for faster analytics and reporting.

1. **Salesforce:** Extracts daily campaign report data
2. **TAP:** Standardizes, validates, and automates data ingestion
3. **/user/SVMDEDMP/hive/salesforce/campaign/campaign\_report\_land/:** Stores raw data in a partitioned structure for efficient querying and processing in Hive.

**Data Processing Pipeline- Salesforce\_Campaign**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svpachds | Vijaya Kumari | 415 |
| svpachdp | Vijaya Kumari | 356 |
| svmdedmp | Hemendra Singh | 266 |
| svfdsroundelhdp | Narendra Yaddula | 14 |
| z00ct7x | Karen Lin | 9 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **03:48:16**

\*Average END Time In CST- **04:02:31**

**\*Average Run Time In CST- 14 Min**

\*Average Start Time In UTC- **09:48:16**

\*Average END Time In UTC- **10:02:31**

**\*Average Run Time In UTC-14 min**

\*Average Throughput-

**2.) Salesforce\_Campaign\_Product- Data Ingestion Process- salesforce → Common api ingestion wrapper →**

**/user/SVMDEDMP/hive/salesforce/campaign\_product/campaign\_product\_report\_daily\_land.**

## **Data Ingestion Process Overview- Salesforce\_Campaign\_Product**

1. **Salesforce:** Extracts daily product-level campaign data
2. **Common API Ingestion Wrapper:** Standardizes, validates, and automates data ingestion
3. **/user/SVMDEDMP/hive/salesforce/campaign\_product/campaign\_product\_report\_daily\_land/:** Stores raw data in a partitioned structure for efficient querying and processing in Hive.

This pipeline ensures consistent, accurate, and scalable integration of **Salesforce Campaign Product Report** data, enabling detailed product-level analytics and data-driven decision-making.

**Data Processing Pipeline- Salesforce\_Campaign\_Products**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svmdedmp | Hemendra Singh | 76 |
| z00ffrs | Anup Jee | 3 |
| z00f1pf | John Deric | 1 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **15:34:30**

\*Average END Time In CST- **16:14:19**

**\*Average Run Time In CST- 40 Min**

\*Average Start Time In UTC- **21:34:30**

\*Average END Time In UTC- **22:14:19**

**\*Average Run Time In UTC-40 min**

\*Average Throughput-

**3.) Salesforce\_Opportunity- Data Ingestion Process- salesforce → Common api ingestion wrapper → /user/SVMDEDMP/hive/salesforce/opportunity/opportunity\_report\_daily\_land.**

## **Data Ingestion Process Overview- Salesforce\_ Opportunity**

* **Salesforce:** Extracts daily opportunity data, including metrics like value, stage, and close dates.
* **Common API Ingestion Wrapper:** Standardizes, validates, and automates the data extraction process, ensuring consistency and accuracy.
* **/user/SVMDEDMP/hive/salesforce/opportunity/opportunity\_report\_daily\_land/:** Stores raw data in a partitioned structure, optimized for Hive queries and historical analysis.

This ingestion pipeline enables consistent, accurate, and scalable integration of **Salesforce Opportunity Report** data It:

1. **Automates Data Extraction:** Ensures timely insights into sales pipeline metrics and opportunity progression.
2. **Maintains Data Quality:** Automated checks for completeness, availability, and accuracy guarantee reliable data.
3. **Optimizes Performance:** Partitioned storage and Parquet format enhance query speed and analytics.
4. **Supports Advanced Analytics:** Facilitates in-depth sales analysis, forecasting, and strategic decision-making.

## **Data Processing Process- Salesforce\_Opportunity**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svfcimip | Jonathan Peters | 349 |
| svrdehdpp | Ami Lathia | 259 |
| svrdehdpd | Ami Lathia | 227 |
| svmdedmp | Hemendra Singh | 87 |
| b032579 | Brian Ohana | 33 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **00:28:22**

\*Average END Time In CST- **00:57:53**

**\*Average Run Time In CST- 30 Min**

Average Start Time In UTC- **06:28:22**

\*Average END Time In UTC- **06:57:53**

**\*Average Run Time In UTC-30 min**

\*Average Throughput-

**4.) Salesforce\_Revenue- Data Ingestion Process- salesforce → Common api ingestion wrapper→/user/SVMDEDMP/hive/salesforce/revenue/revenue\_report\_land**

## **Data Ingestion Process Overview- Salesforce\_ Revenue**

* **Salesforce:** Extracts daily revenue metrics, including bookings, billings, and recognized revenue.
* **Common API Ingestion Wrapper:** Standardizes, validates, and automates the data extraction process, ensuring data consistency and reliability.
* **/user/SVMDEDMP/hive/salesforce/revenue/revenue\_report\_land/:** Stores raw data in a partitioned structure, optimized for Hive queries and historical trend analysis.

This ingestion pipeline enables consistent, accurate, and scalable integration of **Salesforce Revenue Report** data. It:

* **Automates Data Extraction:** Ensures timely insights into revenue metrics, including bookings, billings, and recognized revenue.
* **Maintains Data Quality:** Automated checks for completeness, availability, and accuracy guarantee reliable financial data.
* **Optimizes Performance:** Partitioned storage and Parquet format enhance query speed and analytics.
* **Supports Financial Analysis:** Facilitates in-depth revenue analysis, forecasting, and strategic financial decision-making.

## **Data Processing Process- Salesforce\_Revenue**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svfcimip | Jonathan Peters | 196 |
| svrdehdpp | Ami Lathia | 93 |
| svmdedmp | Hemendra Singh | 83 |
| z00f1pf | John Deric | 2 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th, 2024-January 20th, 2025-none

\*Average Start Time In CST- **22:38:18**

\*Average END Time In CST- **23:10:22**

**\*Average Run Time In CST- 32 Min**

\*Average Start Time In UTC- **04:38:18**

\*Average END Time In UTC- **05:10:22**

**\*Average Run Time In UTC-32 min**

\*Average Throughput-

**5.) Salesforce\_Brand- Data Ingestion Process-** s**alesforce → Common api ingestion wrapper →**

**/user/SVMDEDMP/hive/salesforce/brand/brand\_report\_land**

## **Data Ingestion Process Overview- Salesforce\_ Brand**

This pipeline ensures consistent, accurate, and scalable integration of **Salesforce Brand Report** data, enabling comprehensive brand performance insights and data-driven marketing strategies. This pipeline empowers marketing and analytics teams with comprehensive brand insights, enabling strategic decision-making and improved brand management.

1. **Salesforce:** Extracts daily brand performance data, including metrics like awareness, engagement, and reach.
2. **Common API Ingestion Wrapper:** Standardizes, validates, and automates data extraction, ensuring data consistency and accuracy.
3. **/user/SVMDEDMP/hive/salesforce/brand/brand\_report\_land/:** Stores raw data in a partitioned structure, optimized for Hive queries and historical brand analysis.

## **Data Processing Process- Salesforce\_Brand**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svmdedmp | Hemendra Singh | 84 |
| z00f1pf | John Deric | 1 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **22:30:00**

\*Average END Time In CST- **22:48:47**

**\*Average Run Time In CST- 19 Min**

\*Average Start Time In UTC- **04:30:00**

\*Average END Time In UTC- **04:48:47**

**\*Average Run Time In UTC-19 min**

\*Average Throughput-

**6.) Salesforce\_Brand\_Opportunity- Data Ingestion Process- salesforce → Common api ingestion wrapper →**

**/user/SVMDEDMP/hive/salesforce/brand/brand\_report\_land**

## **Data Ingestion Process Overview- Salesforce\_ Brand\_Opportunity**

### **1. Salesforce**

* **Purpose:** Extracts daily brand opportunity metrics, including pipeline value, opportunity stages, and conversion rates.
* **Data Points:** Captures data such as brand-level opportunity counts, values, and sales progression.
* **Challenge:** Handling large datasets, ensuring data consistency, and maintaining historical records.

### **2. Common API Ingestion Wrapper**

* **Functionality:** Standardizes, validates, and automates data extraction, ensuring consistent API integration.
* **Operations:** Handles authentication (OAuth 2.0), pagination, and rate limiting for seamless data extraction.
* **Data Validation:** Performs checks for completeness, availability, and accuracy before storage.

### **3. Data Lake Path**

**Location:** /user/SVMDEDMP/hive/salesforce/brand\_opportunity/brand\_opportunity\_report\_daily\_land/

* **Purpose:** Stores raw daily brand opportunity data before transformation and validation.
* **Structure:** Organized by date partitions for efficient retrieval and historical analysis.
* **File Format:** Typically stored as **JSON** or **CSV** for raw data, converted to **Parquet** for optimized querying.

## **Data Processing Process- Salesforce\_Brand\_Opportunity**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svmdedmp | Hemendra Singh | 90 |

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **22:30:00**

\*Average END Time In CST- **22:47:02**

**\*Average Run Time In CST- 17 Min**

Average Start Time In UTC- **04:30:00**

\*Average END Time In UTC- **04:47:02**

**\*Average Run Time In UTC-17 min**

\*Average Throughput-

**7.) Salesforce\_Brand\_Agency- Data Ingestion Process- salesforce → Common api ingestion wrapper →**

**/user/SVMDEDMP/hive/salesforce/brand\_agency/brand\_agency\_report\_land.**

## **Data Ingestion Process Overview- Salesforce\_ Brand\_Agency**

### **Data Source: Salesforce**

* **Purpose:** Extracts daily brand-agency metrics, including campaign performance, spend, engagement, and ROI.
* **Data Points:** Captures data such as agency-level spend, impressions, clicks, and conversions for brand campaigns.
* **Challenges:** Handling large datasets, ensuring data consistency, and maintaining historical records.

### **2. Common API Ingestion Wrapper**

* **Functionality:** A reusable module that standardizes, validates, and automates data extraction from Salesforce.
* **Operations:**
  + **Authentication:** Secure access using OAuth 2.0.
  + **Pagination Handling:** Efficiently manage large datasets through paginated API calls.
  + **Rate Limiting & Error Handling:** Handles API rate limits and retries on transient failures.
* **Data Validation:** Checks for completeness, availability, and accuracy before storage.

### **3. Data Landing (Raw Storage)**

* **Location:** /user/SVMDEDMP/hive/salesforce/brand\_agency/brand\_agency\_report\_land/
* **Purpose:** Store raw daily brand-agency data before transformation and validation.
* **Partitioning:** Organized by date for efficient retrieval and historical analysis.
* **File Format:** Typically stored as **JSON** or **CSV** for raw data.

**Data Processing Pipeline- Salesforce\_Brand\_Agency**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| Top Users |  |  |
| LAN ID | Account Owner | Number of Usage Events |
| svmdedmp | Hemendra Singh | 88 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **10:30:00**

\*Average END Time In CST- **10:47:08**

**\*Average Run Time In CST- 17 Min**

\*Average Start Time In UTC- **16:30:00**

\*Average END Time In UTC- **16:47:08**

**\*Average Run Time In UTC-17 min**

\*Average Throughput-

**8.) Salesforce\_Account- Data Ingestion Process-** **salesforce → Common api ingestion wrapper →**

**/user/SVMDEDMP/hive/salesforce/account/account\_report\_land**

## **Data Ingestion Process Overview- Salesforce\_ Account**

### **1. Data Source: Salesforce**

* **Purpose:** Extracts daily brand-agency metrics, including campaign performance, spend, engagement, and ROI.
* **Data Points:** Captures data such as agency-level spend, impressions, clicks, and conversions for brand campaigns.
* **Challenges:** Handling large datasets, ensuring data consistency, and maintaining historical records.

### **2. Common API Ingestion Wrapper**

* **Functionality:** A reusable module that standardizes, validates, and automates data extraction from Salesforce.
* **Operations:**
  + **Authentication:** Secure access using OAuth 2.0.
  + **Pagination Handling:** Efficiently manage large datasets through paginated API calls.
  + **Rate Limiting & Error Handling:** Handles API rate limits and retries on transient failures.
* **Data Validation:** Checks for completeness, availability, and accuracy before storage.

### **3. Data Landing (Raw Storage)**

* **Location:** /user/SVMDEDMP/hive/salesforce/brand\_agency/brand\_agency\_report\_land/
* **Purpose:** Store raw daily brand-agency data before transformation and validation.
* **Partitioning:** Organized by date for efficient retrieval and historical analysis.
* **File Format:** Typically stored as **JSON** or **CSV** for raw data.

**Data Processing Pipeline- Salesforce\_Brand\_Agency**

**Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.

|  |  |  |
| --- | --- | --- |
| LAN ID | Account Owner | Number of Usage Events |
| svrikbrd | L Shekhar | 295 |
| svfcimip | Jonathan Peters | 198 |
| svmdedmp | Hemendra Singh | 89 |
| svrdehdpp | Ami Lathia | 85 |
| svrpihds | L Shekhar | 85 |

**Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.

Failure Rate- Dated October 20th 2024-January 20th 2025-none

\*Average Start Time In CST- **10:30:00**

\*Average END Time In CST- **10:47:08**

**\*Average Run Time In CST- 17 Min**

\*Average Start Time In UTC- **16:30:00**

\*Average END Time In UTC- **16:47:08**

**\*Average Run Time In UTC-17 min**

\*Average Throughput-

**9.) Salesforce\_Class- Data Ingestion Process-**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **salesforce → Common api ingestion wrapper →**  **/user/SVMDEDMP/hive/salesforce/class/class\_report\_daily\_land** **1. Data Source: Salesforce**  * **Purpose:** Extracts daily account data, including account name, industry, revenue, account owner, and status. * **Data Points:** Captures key account information such as contact details, account type, and custom fields. * **Challenges:** Handling large datasets, ensuring data consistency, and managing incremental updates.  **2. Common API Ingestion Wrapper**  * **Functionality:** A reusable module that standardizes, validates, and automates data extraction from Salesforce. * **Operations:**   + **Authentication:** Secure access using OAuth 2.0.   + **Pagination Handling:** Efficiently manage large datasets through paginated API calls.   + **Incremental Data Extraction:** Captures only new and updated records for optimized processing.   + **Error Handling:** Retries on transient API failures and logs errors for auditing. * **Data Validation:** Checks for completeness, accuracy, and uniqueness before storage.  **3. Data Landing (Raw Storage)**  * **Location:** /user/SVMDEDMP/hive/salesforce/account/account\_report\_land/ * **Purpose:** Stores raw daily account data before transformation and validation. * **Partitioning:** Organized by date for efficient retrieval and historical analysis. * **File Format:** Typically stored as **JSON** or **CSV** for raw data dumps.   **Data Processing Pipeline- Salesforce\_Class**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svfcimip | Jonathan Peters | 180 | | svrdehdpp | Ami Lathia | 87 | | svmdedmp | Hemendra Singh | 85 | | svmdedmd | Hemendra Singh | 2 | | b032579 | Brian Ohana | 1 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **10:30:00**  \*Average END Time In CST- **10:47:04**  **\*Average Run Time In CST- 17 Min**  \*Average Start Time In UTC- **16:30:00**  \*Average END Time In UTC- **16:47:04**  **\*Average Run Time In UTC-17 min**  \*Average Throughput-  **10.) Salesforce\_Contact- Data Ingestion Process-** **salesforce → Common api ingestion wrapper →**  **/user/SVMDEDMP/hive/salesforce/contact/contact\_report\_land** **Data Source: Salesforce**  * **Purpose:** Extracts daily contact data, including names, emails, phone numbers, titles, and account associations. * **Data Points:** Captures key contact details such as department, status, and custom fields. * **Challenges:** Handling large datasets, ensuring data consistency, and managing incremental updates.  **2. Common API Ingestion Wrapper**  * **Functionality:** A reusable module that standardizes, validates, and automates data extraction from Salesforce. * **Operations:**   + **Authentication:** Secure access using OAuth 2.0.   + **Pagination Handling:** Efficiently manage large datasets through paginated API calls.   + **Incremental Data Extraction:** Captures only new and updated records for optimized processing.   + **Error Handling:** Retries on transient API failures and logs errors for auditing. * **Data Validation:** Checks for completeness, accuracy, and uniqueness before storage.  **3. Data Landing (Raw Storage)**  * **Location:** /user/SVMDEDMP/hive/salesforce/contact/contact\_report\_land/ * **Purpose:** Stores raw daily contact data before transformation and validation. * **Partitioning:** Organized by date for efficient retrieval and historical analysis. * **File Format:** Typically stored as **JSON** or **CSV** for raw data dumps.   **Data Processing Pipeline- Salesforce\_Contact**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svmdedmp | Hemendra Singh | 85 | | b032579 | Brian Ohana | 5 | | z006ypl | Ben Brochtrup | 1 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **10:30:00**  \*Average END Time In CST- **10:47:04**  **\*Average Run Time In CST- 17 Min**  \*Average Start Time In UTC- **16:30:00**  \*Average END Time In UTC- **16:47:04**  **\*Average Run Time In UTC-17 min**  \*Average Throughput-  **11.) Salesforce\_Vendor- Data Ingestion Process- salesforce → Common api ingestion wrapper →**  **/user/SVMDEDMP/hive/salesforce/vendor/vendor\_report\_land** **1. Data Source: Salesforce**  * **Purpose:** Extracts daily vendor data, including vendor names, contact details, contracts, payment terms, and performance metrics. * **Data Points:** Captures key vendor information such as industry, status, spend, and custom fields. * **Challenges:** Handling large datasets, ensuring data consistency, and managing incremental updates.  **2. Common API Ingestion Wrapper**  * **Functionality:** A reusable module that standardizes, validates, and automates data extraction from Salesforce. * **Operations:**   + **Authentication:** Secure access using OAuth 2.0.   + **Pagination Handling:** Efficiently manage large datasets through paginated API calls.   + **Incremental Data Extraction:** Captures only new and updated records for optimized processing.   + **Error Handling:** Retries on transient API failures and logs errors for auditing. * **Data Validation:** Checks for completeness, accuracy, and uniqueness before storage.  **3. Data Landing (Raw Storage)**  * **Location:** /user/SVMDEDMP/hive/salesforce/vendor/vendor\_report\_land/ * **Purpose:** Stores raw daily vendor data before transformation and validation. * **Partitioning:** Organized by date for efficient retrieval and historical analysis. * **File Format:** Typically stored as **JSON** or **CSV** for raw data dumps.   **Data Processing Pipeline- Salesforce\_Contact**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svrikbrd | L Shekhar | 93 | | svfcimip | Jonathan Peters | 88 | | svmdedmp | Hemendra Singh | 86 | | svrpihds | L Shekhar | 84 | | svrdehdpp | Ami Lathia | 83 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **10:30:00**  \*Average END Time In CST- **10:46:37**  **\*Average Run Time In CST- 17 Min**  Average Start Time In UTC- **16:30:00**  \*Average END Time In UTC- **16:46:37**  **\*Average Run Time In UTC-17 min**  \*Average Throughput-  **12.) Salesforce\_User- Data Ingestion Process- salesforce → Common api ingestion wrapper →**  **/user/SVMDEDMP/hive/salesforce/user/user\_report\_land** **Data Source: Salesforce**  * **Purpose:** Extracts daily user data, including user names, roles, email addresses, login activity, and permissions. * **Data Points:** Captures key user information such as status (active/inactive), profiles, and custom fields. * **Challenges:** Handling large datasets, ensuring data consistency, and managing incremental updates.  **2. Common API Ingestion Wrapper**  * **Functionality:** A reusable module that standardizes, validates, and automates data extraction from Salesforce. * **Operations:**   + **Authentication:** Secure access using OAuth 2.0.   + **Pagination Handling:** Efficiently manage large datasets through paginated API calls.   + **Incremental Data Extraction:** Captures only new and updated records for optimized processing.   + **Error Handling:** Retries on transient API failures and logs errors for auditing. * **Data Validation:** Checks for completeness, accuracy, and uniqueness before storage.  **3. Data Landing (Raw Storage)**  * **Location:** /user/SVMDEDMP/hive/salesforce/user/user\_report\_land/ * **Purpose:** Stores raw daily user data before transformation and validation. * **Partitioning:** Organized by date for efficient retrieval and historical analysis. * **File Format:** Typically stored as **JSON** or **CSV** for raw data dumps.   **Data Processing Pipeline- Salesforce\_User**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svrikbrd | L Shekhar | 280 | | svrpihds | L Shekhar | 162 | | svrpihdp | L Shekhar | 113 | | svmdedmp | Hemendra Singh | 88 | | svrdehdpp | Ami Lathia | 86 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **10:30:00**  \*Average END Time In CST- **10:46:48**  **\*Average Run Time In CST- 17 Min**  \*Average Start Time In UTC- **16:30:00**  \*Average END Time In UTC- **16:46:48**  **\*Average Run Time In UTC-17 min**  \*Average Throughput-  **4.Criteo-**  **List Of Tables Under Criteo: -**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Schema Name** | **Table Name** | **SLA** | **Load\_Time** | **Load\_Frequency** | **frequecy\_of\_ingestion\_with\_time** | **Ingestion\_flow** | | prd\_mdf\_fnd | criteo\_campaign\_dly | 11.30PM CST | 12.30AM CST | Daily | 1, 12.30AM CST | criteo→ sftp server → sb → s3poller → /user/SVMDEDMP/hive/criteo/criteo\_campaign\_dly/criteo\_campaign\_dly\_land | |  | criteo\_category\_dly | 11AM CST | 12.30AM CST | Daily | 1, 12.30AM CST | criteo→ sftp server → sb → s3poller → /user/SVMDEDMP/hive/criteo/criteo\_category\_dly/criteo\_category\_dly\_land | |  | criteo\_category\_advertiser\_dly | 11AM CST | 12.30AM CST | Daily | 1, 12.30AM CST | criteo→ sftp server → sb → s3poller → /user/SVMDEDMP/hive/criteo/criteo\_category\_advertiser\_dly/criteo\_category\_advertiser\_dly\_land | |  | criteo\_category\_keyword\_dly | 10.30PM CST | 12.30AM CST | Daily | 1, 12.30AM CST | criteo→ sftp server → sb → s3poller → /user/SVMDEDMP/hive/criteo/criteo\_category\_keyword\_dly/criteo\_category\_keyword\_dly\_land | |  | criteo\_monthly | NA | 12.30AM CST | Daily | 1, 12.30AM CST |  | |  | criteo\_floor\_price | NA | NA | NA | NA | NA | |  | criteo\_capout\_daily | 2:00 PM CST | |  | | --- | | At 12:40 CST | | Daily | 1, At 12:40 CST | criteo->file express→ common ingestion ->/user/SVMDEDMP/hive/criteo/criteo\_capout\_dly/criteo\_capout\_dly\_land | |  | criteo\_dsp\_category\_advertiser\_dly | At 11:00 AM | At 05:30 UTC | Daily | 1, At 05:30 UTC | criteo->file express→ common ingestion →/user/SVMDEDMP/hive/criteo\_dsp/criteo\_dsp\_keyword\_advertiser\_dly/daily\_landing | |  | criteo\_dsp\_category\_keyword\_dly | At 11:00 AM | At 05:30 UTC | Daily | 1, At 05:30 UTC | criteo->file express→ common ingestion →/user/SVMDEDMP/hive/criteo\_dsp/criteo\_dsp\_keyword\_advertiser\_dly/daily\_landing |  **Data Ingestion Process Overview-Criteo-****Similarities Between All Ingestion Pipelines**  1. **Source: Criteo**    1. All pipelines begin with data from **Criteo**, capturing various metrics related to campaigns, categories, advertisers, keywords, and DSP performance. 2. **Data Transfer Mechanisms**    1. Most pipelines use **SFTP Server → SB (Service Bus) → S3Poller** for secure data transfer, while some use **File Express → Common Ingestion** for faster, standardized ingestion.    2. Consistent scheduling and automation using workflow orchestrators like **Apache Airflow** or **Oozie**. 3. **Data Landing & Storage Structure**    1. All land data in the **Hive-based data lake** within the user directory: /user/SVMDEDMP/hive/criteo/.    2. Organized under daily landing zones (\_dly\_land) for efficient partitioning and historical analysis. 4. **Automation & Orchestration**    1. All pipelines are automated using DAG configurations:       1. Data Pull → Data Transfer (SFTP/SB/File Express) → Land Raw Data → Validation → Monitoring. 5. **Data Validation & Quality Checks**    1. All pipelines implement **completeness**, **availability**, and **accuracy** checks before storage.    2. Use consistent logging and alert mechanisms for data quality monitoring. 6. **Security & Access Control**    1. Same security configurations using **Apache Ranger** or **Sentry**.    2. Role-based access control and data masking for sensitive metrics.  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Aspect** | **Campaign Dly** | **Category Dly** | **Category Advertiser Dly** | **Category Keyword Dly** | **Capout Dly** | **DSP Keyword Advertiser Dly** | | **Data Type** | Campaign-level metrics (e.g., impressions, clicks, spend) | Category-level metrics (e.g., performance by product category) | Advertiser-level metrics within categories | Keyword-level metrics within categories | Capout metrics (e.g., budget caps, pacing limits) | DSP performance metrics by keyword and advertiser | | **Landing Location** | /user/SVMDEDMP/hive/criteo/criteo\_campaign\_dly/criteo\_campaign\_dly\_land | /user/SVMDEDMP/hive/criteo/criteo\_category\_dly/criteo\_category\_dly\_land | /user/SVMDEDMP/hive/criteo/criteo\_category\_advertiser\_dly/criteo\_category\_advertiser\_dly\_land | /user/SVMDEDMP/hive/criteo/criteo\_category\_keyword\_dly/criteo\_category\_keyword\_dly\_land | /user/SVMDEDMP/hive/criteo/criteo\_capout\_dly/criteo\_capout\_dly\_land | /user/SVMDEDMP/hive/criteo\_dsp/criteo\_dsp\_keyword\_advertiser\_dly/daily\_landing | | **Transfer Mechanism** | SFTP Server → SB → S3Poller | SFTP Server → SB → S3Poller | SFTP Server → SB → S3Poller | SFTP Server → SB → S3Poller | File Express → Common Ingestion | File Express → Common Ingestion | | **Partitioning Strategy** | Partitioned by **campaign\_id** and **date** | Partitioned by **category\_id** and **date** | Partitioned by **advertiser\_id**, **category\_id**, and **date** | Partitioned by **keyword\_id**, **category\_id**, and **date** | Partitioned by **capout\_type** and **date** | Partitioned by **keyword\_id**, **advertiser\_id**, and **date** | | **Data Model Structure** | Captures campaign attributes like campaign name, budget, and targeting | Captures category attributes like category name, product count, and performance | Captures advertiser-specific metrics within categories | Captures keyword-level performance within categories | Captures budget cap, pacing limits, and delivery constraints | Captures DSP performance by keyword and advertiser | | **Aggregation & Reporting** | Aggregates data by **campaign** for performance analysis | Aggregates data by **category** for product segmentation analysis | Aggregates data by **advertiser** within categories | Aggregates data by **keyword** within categories | Aggregates capout data for budget management | Aggregates DSP metrics for keyword and advertiser performance | | **Use Cases** | Used for campaign performance tracking, optimization, and ROI analysis | Used for category performance insights, inventory planning, and marketing strategy | Used for advertiser-level insights within categories | Used for keyword-level insights within categories | Used for budget management, pacing, and delivery optimization | Used for DSP keyword and advertiser performance analysis | | **Hive Table Definitions** | External table with campaign-specific schema | External table with category-specific schema | External table with category-advertiser schema | External table with category-keyword schema | External table with capout-specific schema | External table with DSP keyword-advertiser schema | | **Query Patterns** | Queries focus on campaign-level KPIs (e.g., CPC, CPA) | Queries focus on category-level KPIs (e.g., conversion rate by category) | Queries focus on advertiser-level KPIs within categories | Queries focus on keyword-level KPIs within categories | Queries focus on budget utilization and pacing metrics | Queries focus on DSP performance by keyword and advertiser |  **Key Differences**  1. **Data Type & Granularity**    1. Pipelines vary in data granularity, ranging from **campaign-level** to **keyword-advertiser level**.    2. **Capout** and **DSP Keyword Advertiser** pipelines capture specialized metrics for budget management and DSP performance. 2. **Landing Location & Partitioning Strategy**    1. Different landing locations in Hive, partitioned by unique identifiers (e.g., campaign\_id, category\_id, advertiser\_id, keyword\_id). 3. **Transfer Mechanism**    1. Most pipelines use **SFTP Server → SB → S3Poller**, but **Capout** and **DSP Keyword Advertiser** use **File Express → Common Ingestion** for standardized ingestion. 4. **Data Model Structure & Use Cases**    1. Each pipeline caters to specific business needs, from campaign performance analysis to DSP optimization and budget management.   **Data Processing Pipeline- Criteo-**  The processing Process Differs for all the tables mentioned above-  **1.) criteo\_campaign\_dly-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svrpihds | L Shekhar | 649 | | svrmshdp | Matthew Christianson | 211 | | svcylhds | Nomvelo Moyo | 163 | | svrpihdp | L Shekhar | 96 | | svmdedmp | Hemendra Singh | 90 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **12:19:35**  \*Average END Time In CST- **13:57:53**  **\*Average Run Time In CST- 98 Min**  \*Average Start Time In UTC- **18:19:35**  \*Average END Time In UTC- **19:57:53**  **\*Average Run Time In UTC-98 min**  \*Average Throughput-  **2.) criteo\_category\_dly-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svmdedmp | Hemendra Singh | 82 | | svmdedmd | Hemendra Singh | 20 | | svrdehdpp | Ami Lathia | 14 | | z003lwr | Rajee Gupta | 8 | | z00ffrs | Anup Jee | 6 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **06:46:38**  \*Average END Time In CST- **07:25:42**  **\*Average Run Time In CST- 39 Min**  \*Average Start Time In UTC- **12:46:38**  \*Average END Time In UTC- **13:25:42**  **\*Average Run Time In UTC-39 min**  \*Average Throughput-  **3.) criteo\_category\_advertiser\_dly-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svfcimip | Jonathan Peters | 7 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- None  \*Average END Time In CST- None  **\*Average Run Time In CST-** None  \*Average Start Time In UTC- None  \*Average END Time In UTC- None  **\*Average Run Time In UTC-** None  \*Average Throughput-  **4.) criteo\_category\_Keyword\_dly-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svmdedmp | Hemendra Singh | 83 | | z00dxyh | Shraddha Vimal | 39 | | z00f1wz | Anil Javvadula | 11 | | z008zzf | Sourav Saha | 7 | | svmdedmd | Hemendra Singh | 6 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **11:35:54**  \*Average END Time In CST- **14:21:32**  **\*Average Run Time In CST- 166 min**  \*Average Start Time In UTC- **17:35:54**  \*Average END Time In UTC- 20**:21:32**  **\*Average Run Time In UTC-** 166 min  \*Average Throughput-  **5.) Criteo\_Monthly-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svrpihds | L Shekhar | 738 | | svrdehdpt | Ami Lathia | 495 | | svrdehdpd | Ami Lathia | 231 | | svrmshdp | Matthew Christianson | 211 | | svmdedmd | Hemendra Singh | 166 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **03:19:02**  \*Average END Time In CST- **04:39:58**  **\*Average Run Time In CST- 81 min**  \*Average Start Time In UTC- **09:19:02**  \*Average END Time In UTC- **10:39:58**  **\*Average Run Time In UTC-** 81 min  \*Average Throughput-  **6.) Criteo\_Floor\_Price-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svmdedmp | Hemendra Singh | 3 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- None  \*Average END Time In CST- None  **\*Average Run Time In CST-** None  \*Average Start Time In UTC- None  \*Average END Time In UTC- None  **\*Average Run Time In UTC-** None  \*Average Throughput-  **7.) Criteo\_Capout\_Daily-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svrpihds | L Shekhar | 1118 | | svrpihdp | L Shekhar | 321 | | svmdedmp | Hemendra Singh | 93 | | svmdedmd | Hemendra Singh | 4 | | z001kh3 | Leo Prince Francis Xavier | 3 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **02:27:02**  \*Average END Time In CST- **02:37:35**  **\*Average Run Time In CST- 11 min**  \*Average Start Time In UTC- **08:27:02**  \*Average END Time In UTC- **08:37:35**  **\*Average Run Time In UTC-** 11 min  \*Average Throughput-  **8.) Criteo\_Dsp\_Category\_Advertiser\_Dly-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svrpihds | L Shekhar | 471 | | svrpihdp | L Shekhar | 278 | | svmdedmp | Hemendra Singh | 150 | | z006yql | Rahul Bindra | 22 | | svfdsroundelhdp | Narendra Yaddula | 7 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **11:42:02**  \*Average END Time In CST- **14:03:16**  **\*Average Run Time In CST- 141 min**  \*Average Start Time In UTC- **17:42:02**  \*Average END Time In UTC- **20:03:16**  **\*Average Run Time In UTC-** 141 min  \*Average Throughput-  **9.) Criteo\_Dsp\_Category\_Keyword\_Dly-**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svmdedmp | Hemendra Singh | 157 | | z00dxyf | RAHUL MITTAL | 68 | | svmdedmd | Hemendra Singh | 16 | | z00ffrs | Anup Jee | 15 | | z006yql | Rahul Bindra | 8 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST- **11:42:02**  \*Average END Time In CST- **14:03:16**  **\*Average Run Time In CST- 141 min**  \*Average Start Time In UTC- **17:42:02**  \*Average END Time In UTC- **20:03:16**  **\*Average Run Time In UTC-** 141 min  \*Average Throughput-  **5.Citrus-**  **List Of Tables Under Citrus:**   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Schema Name** | **Table Name** | **SLA** | **Load\_Time** | **Load\_Frequency** | **frequecy\_of\_ingestion\_with\_time** | **Ingestion\_flow** | | prd\_mdf\_fnd | citrus\_api\_actuals | 5:00 PM | |  | | --- | | 12:00 AM | | Daily | trigger file | API→**switchboard**→S3→TAP app->hdfs |  **Data Ingestion Process Overview-Citrus-** This ingestion pipeline is designed to extract data from an **API**, process and route it through **Switchboard**, temporarily store it in **Amazon S3**, and then move it into **HDFS (Hadoop Distributed File System)** using **TAP App**. It ensures:   * **Automated Data Extraction:** Captures data from APIs with efficient handling of pagination and rate limits. * **Scalable Storage & Transfer:** Utilizes S3 for scalable, temporary storage and reliable data transfer. * **Standardized Ingestion:** Uses TAP App for consistent transformation and loading into HDFS. * **Optimized Performance:** Efficient data partitioning and storage in HDFS for analytics and processing.  **1. Source: API**  * **Purpose:** Extracts data from a third-party or internal API, including metrics, events, or transactional data. * **Challenges:** Handling authentication, pagination, rate limiting, and ensuring data consistency. * **Solution:** Using an API integration module to:   + **Authentication:** Secure access using OAuth 2.0 or API keys.   + **Pagination Handling:** Efficiently manage large datasets through paginated API calls.   + **Error Handling:** Retries on transient API failures and logs errors for auditing.  **2. Switchboard**  * **Functionality:** A data routing and transformation layer for flexible data flow management. * **Operations:**   + **Data Routing:** Routes API data to the appropriate S3 bucket based on source and data type.   + **Transformation:** Basic transformations such as data normalization, filtering, and validation.   + **Data Enrichment:** Optional enrichment by joining with other data sources before landing in S3.   + **Monitoring & Alerts:** Real-time monitoring and alerts for data transfer failures or delays.  **3. S3 (Amazon Simple Storage Service)**  * **Purpose:** Temporary storage for raw API data before further processing and loading into HDFS. * **Structure:** Organized by source, data type, and date partitions for efficient retrieval. * **Example Directory Structure:**   bash  Copyedit  s3://bucket-name/source\_name/data\_type/YYYY/MM/DD/raw\_data.json s3://bucket-name/api\_name/events/2025/02/17/raw\_events.json   * **File Format:** Typically stored as **JSON** or **CSV** for raw data. * **Benefits:** Scalable, cost-effective storage with high durability and availability.  **4. TAP App (Transfer and Processing Application)**  * **Functionality:** Standardizes, validates, and automates data extraction from S3 to HDFS. * **Operations:**   + **Data Extraction:** Polls S3 buckets for new files using event triggers or scheduled jobs.   + **Data Validation:** Checks for completeness, availability, and accuracy before loading into HDFS.   + **Transformation:** Normalizes and standardizes data formats for consistent storage.   + **Error Handling:** Retries failures and logs errors for auditing and monitoring.   + **Monitoring & Alerts:** Real-time monitoring and alerts for data quality breaches or failures.  **5. HDFS (Hadoop Distributed File System)**  * **Location:** /user/SVMDEDMP/hive/source\_name/data\_type/landing\_zone/ * **Purpose:** Stores transformed data in a partitioned structure optimized for processing and analytics. * **Partitioning:** Organized by date for efficient retrieval and historical analysis. * **File Format:** Typically stored as **Parquet** or **ORC** for optimized querying.   **Data Processing Pipeline- Citrus: -**  **Usage**- Refers to the users who utilize the tables to extract data within the specified time frame. (Duration-90 days)-Dated October 20th, 2024-January 20th, 2025.   |  |  |  | | --- | --- | --- | | LAN ID | Account Owner | Number of Usage Events | | svrdehdpt | Ami Lathia | 237 | | svdip001p | Darian Hall | 219 | | svrdehdpp | Ami Lathia | 157 | | svmdedmp | Hemendra Singh | 83 | | svtgtphdp | SHOBANA S | 14 |   **Lineage**- In a data portal, **data lineage** refers to the process of tracking and visualizing the flow of data through its lifecycle, from its origin to its destination. It provides insight into how data is created, transformed, and consumed, helping users understand the data’s journey and the relationships between different data elements.  Failure Rate- Dated October 20th 2024-January 20th 2025-none  \*Average Start Time In CST-**11:02:57**  \*Average END Time In CST- **11:56:29**  **\*Average Run Time In CST- 54 min**  \*Average Start Time In UTC- **17:02:57**  \*Average END Time In UTC- **17:56:29**  **\*Average Run Time In UTC-** 54 min  \*Average Throughput   |  | | --- | |  |  |  | | --- | |  | |