PRASATH SRINIVASAN

- :prasathsrinivasan07@gmail.com :+91 9952390771

Professional Summary

- A collaborative engineering professional with substantial knowledge and **experience for about 4 Years,** in **Analysis, Design, Development, Implementation, migration, convergence, Manage, implement and Support** of large **DBs, Data-warehouse and Big Data** systems by creating an intuitive **architectures** and **frameworks** that helps organizations effectively **capture, store, process and analyze huge volume of structured, semi structured, unstructured** and stream of heterogeneous data set using batch and In memory techniques Extensive working experience on Hadoop and its components like **Hdfs, Mapreduce, Sqoop, SQL**, **Hive** and **Spark**.
- Benchmarking Hive using different file format storage like **Avro, orc, parquet**. Handling the evolving schema (**Hive Dynamic Schema**).
- Good understanding of storage formats (**Ser-De**) and Optimization Techniques like **Push Down optimization**, **Compression Techniques**, **Join Optimization** and **Cost Based Optimization** (CBO).
- Good understanding on both **Partitions, Bucketing** and designed both **Managed** and **External** tables in Hive to optimize performance. Having good understanding of storage level of cache and persist.
- Very good understanding on **Spark Performance tuning** in **Pushdown Optimizations** (**Predicate, Projection** and **Partition pruning**) and **Spark SQL joining** (**Broadcast, SMB join, shuffle hash join and Skew join**)
- Have a good understanding of the **GCP Dataproc** to create cluster with the feature of **Cost saving** by applying **Start/Stop** (**Long Available cluster**) and **Ephemeral cluster**.
- Extensive working experience on **GCP** and its components **Dataproc, GCS**, **Big Query, Cloud Composer** and **Airflow**.
- Extensive working experience on **GCP** and its components **Pub/sub and Dataflow.**

SKILLS

- Hadoop
- HDFS
- Sqoop
- Map-Reduce
- Yarn
- Spark
- Python
- Pyspark
- SQL
- Kafka
- Nifi
- GCP
- GCS
- IAM
- EMR
- Dataproc
- Big query
- Data prep
- Data flow
- Pub/sub
- Cloud Composer
- Airflow
- Linux
- CI/CD

EDUCATION

Master of Engineering, Engineering Design
St. Peter's University, Chennai –India

LANGUAGES

Tamil, English, Telugu

PROJECT 2

Title: Customer Segmentation Data Lake and Analytics
Client: Charter Communications Inc.
Period: April-2022 till Now
Organization: Inceptez Technologies Private Limited, Chennai
Technology: SQL, Python, PySpark, Dataproc, GCS, Bigquery, Cloud Composer
Project Description: Retailers can analyze this data to generate insights about individual consumer behaviors and preferences, Recommendations and Loyalty Programs. This tool builds the strategies, Key Performance Indicators (KPI) definitions and implementation roadmaps that assists our esteemed clients in their Analytics & Information ecosystem journey right from Strategy definition to large scale Global Implementations & Support using the Big data ecosystems.
Responsibilities:

- Cron -> Shell Script -> data availability -> Start Cluster -> Submit Spark job -> load BQ tables -> generate BQ reports (can happen any time even cluster is stopped) -> Stop Cluster
- Create a Cloud Composer DAG (Orchestration) and schedule the Workflow to run
- By creating views (Redaction/Masking) the data can be certified to use further for applicable tables.
- Will do a Data Exploration/Data Discovery (EDA) -> Data Munging (Cleansing/Scrubbing) -> Transformation(Curation) -> Data enrichment
- Data Structurizations/Standardization/schema migration (nested data to simple data using unnest function)
- Create temp table for further downstream
- Apply wrangling, munging transformations
- Applying Performance Improvements: Use native tables, use partitions and clusters, use temp tables rather using CTE/inline views to avoid memory errors, use preview option rather using limit, avoid using auto schema, use truncate rather using delete, use delete rather using merge update.

PROJECT 1

Title: Define the Insurance Premium Based on Credit card Defaulters
Client: ING Vysya
Period: July-2019 to Jan-2022
Organization: Inceptez Technologies Private Limited, Chennai
Technology: SQL, HDFS, S3, Shell Script, HIVE
Project Description: This project helps the Financial institutions such and Insurance and Banking systems understands the Credit card defaulters to define the Insurance premium accordingly, it analyze a wide variety of online and offline customer data including the customer transactions, customer master data, insurance data etc.
Responsibilities:

- (RDBMS,LFS,S3)-> Shell Script -> Sqoop -> HDFS -> Hive Parser -> Hive -> RDBMS
- Merge the 2 dataset using hive and split the defaulters and non-defaulters into 2 data sets and load into hdfs.
- Create and load penalty data into the hive table serialized with orc.
- Partitioning of different parameters by suitable method done by Hive.
- Export the Defaulters and Non Defaulters data into HDFS with comma delimiter, where non defaulter's data added with trailer data for consumer systems validation.
- Apply Data Governance - Redaction and Masking using Hive and Python
- Write queries to build cubes at different levels to aggregate the data in real time to populate in the report such as average age, sum of bill amount, average bill amount etc.
- Applying Performance Improvements: Hive partitioning it distributes execution load horizontally, faster execution of queries, Bucketing is allow to decompose data, serialization and deserialization allows to create custom SerDes based on the data definition.