**SAI KRISHNA**

**Sr. Big Data Engineer**

**Summary**

* 11+ years of Experience in **Data Analysis, Data Modeling and BigData** professional with applied information Technology.
* Experienced in Technical consulting and end-to-end delivery with **dataanalysis, data modeling, data governance** and design - development - implementation of solutions.
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to Google Cloud Platform(GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Composer.
* Have experience in **Apache Spark, Spark Streaming, Spark SQL** and **NoSQL** databases like **HBase, Cassandra**, and **MongoDB.**
* Experience in developing Map Reduce Programs using **Apache Hadoop** for analyzing the big data as per the requirement.
  + Good Knowledge with cloud technologies like **Azure** and **AWS , Databricks**(EMR, S3, Red Shift, EC2, DynamoDB).
  + Worked with **NoSQL** databases like **HBase, Cassandra** and **MongoDB** for information extraction and place huge amount of data.
  + Expertise in Big Data Ingestion/Integration Tool like **flume, Kafka**.
  + Very keen in knowing never techno stack that Google Cloud Platform (GCP) adds.
* Practical understanding of the Data modeling (Dimensional & Relational) concepts like **Star-Schema Modeling**, **Snowflake Schema** Modeling, Fact and Dimension tables.
* Good knowledge in using apache **NiFi** to automate the data movement between different Hadoop systems.
  + Used **Informatica Power Center** for Extraction, Transformation, and Loading (ETL) of information from numerous sources like Flat files, XML documents, and Databases.
* Expertise in **Data Governance, Collibra Software**, and Business Analytics.
  + Hands on experience in **Normalization** (1NF, 2NF, 3NF and BCNF) **Denormalization** techniques for effective and optimum performance in **OLTP** and **OLAP** environments.
  + Strong background in various Data Modeling tools using **ERWIN, ER Studio, MS Visio**.
  + Extensive experience in **Relational Data Modeling**, Dimensional Data Modeling, Logical data model/**Physical data models** Designs, **ER Diagrams**, Forward and Reverse Engineering, Publishing **ERWIN diagrams**, analyzing data sources and creating interface documents.
  + Designed and developed Data Marts by following **Star Schema** and Snowflake Schema Methodology, using industry leading Data Modeling tools like **ERWin.**
  + Experience on **Palantir** Foundryand Data warehouses (SQL Azure and Confidential Redshift/RDS).
  + Solid knowledge of Data Marts, Operational Data Store **(ODS),** OLAP, Dimensional Data Modeling with Ralph Kimball Methodology (Star Schema Modeling, **Snow-Flake Modeling** for FACT and Dimensions Tables) using Analysis Services.
  + Expertise in Developing Big data solutions using Data ingestion, Data Storage.
  + Experience in cluster monitoring tools like **Apache hue.**
  + Have 2+ years of hands on experience with Snowflake DBMS.
  + Good experience in using **Sqoop** for traditional RDBMS data pulls.
  + Spark for ETL follower. Databricks Enthusiast, Cloud Adoption & Data Engineering enthusiast in Open source community.
  + Strong experience in database skills in IBM- DB2, Oracle and Proficient in database development, including Constraints, Indexes, **Views, Stored Procedures, Triggers** and Cursors.
  + Proficient with Azure Data Lake Services (ADLS), Databricks & iPython Notebooks formats, Databricks Deltalakes & Amazon Web Services (AWS).
  + Strong understanding of Data Modelling (Relational, dimensional, Star and Snowflake Schema), Data analysis, **Palantir** Foundry, implementations of Data warehousing using Windows and UNIX.
  + Extensive use of Open Source Software and Web/Application Servers like Eclipse 3.x IDE and **Apache Tomcat 6.0.**
  + Experience in designing a component using UML Design-**Use Case, Class, Sequence**, and Development, Component diagrams for the requirements.
  + Excellent knowledge of Machine Learning, Mathematical Modeling and Operations Research. Comfortable with R, **Python**, SAS and Weka, MATLAB, Relational databases. Deep understanding & exposure of Big Data Eco - system.
  + Hands on experience on developing **UDF, DATA Frames** and SQL Queries in **Spark SQL.**
  + Highly skilled in integrating **Kafka** with Spark streaming for high speed data processing.
  + Understanding of data storage and retrieval techniques, ETL, and databases, to include graph stores, relational databases, tuple stores.
* Logical and physical database designing like Tables, Constraints, Index, etc. using Erwin, **ER Studio, TOAD Modeler** and **SQL Modeler**.
  + Experienced in writing Storm topology to accept the events from Kafka producer and emit into **Cassandra DB.**
  + Excellent knowledge in SQL and coding PL/SQL Packages, **Procedures.**
  + Capable at using AWS utilities such as **EMR, S3** and Cloud watch to run and monitor **Hadoop/Spark** jobs on AWS.
  + Good understanding and exposure to **Python** programming.
  + Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
  + Developed PL/SQL programs (**Functions,** Procedures, Packages and **Triggers**).
  + Involved in reports development using reporting tools like **Tableau**. Used excel sheet, flat files, **CSV** files to generated Tableau adhoc reports.
  + Broad design, development and testing experience with **Talend Integration Suite** and knowledge in Performance tuning of mappings.

**Technical Skills**

**Hadoop/Big Data:** HDFS, MapReduce, Hive, Pig, Sqoop, Knime,Flume, Oozie, Spark, Kafka, Storm and Zookeeper, NiFi.

**Data Modeling Tools:** Erwin, Oracle Designer, ER/Studio.

**ETL Tools:** Pentaho, Informatica Power 9.6 etc.

**Operating Systems**:HP-UNIX, RedHat Linux, Ubuntu Linux and Windows.

**Cloud Platform:** Azure, AWS.

**Databases:** Oracle 12c/11g, Teradata R15/R14, MS SQL Server 2016/2014, DB2.

**OLAP Tools**: Tableau 7, SAP BO, SSAS, Business Objects, and Crystal Reports 9

**No SQL Databases** : HBase, Cassandra, MongoDB.

**Web/Application servers**:Apache Tomcat, WebLogic, JBoss.

**Tools and IDE**:Eclipse, NetBeans, Toad, Maven, ANT, Hudson, Sonar, JDeveloper, Assent PMD, DB Visualizer.

**Version control:** SVN, CVS, GIT.

**Web Services**:REST, SOAP.

**Languages:** C, Python, Scala, PL/SQL, Pig Latin, HiveQL, Unix shell scripts.

**Professional Summary**

**Client: Drop box, San Francisco, CA Oct 2018 to June 2023 Role: Sr. Big Data Engineer**

**Responsibilities:**

* Analyzed large and critical datasets using **HDFS, HBase, MapReduce, , Hive**, Hive UDF, **Pig, Sqoop, Zookeeper** and **Spark.**
* Loaded and transformed large sets of structured, semi structured and unstructured data using **Hadoop**/Big Data concepts.
* Performed Data transformations in **HIVE** and used partitions, buckets for performance improvements.
  + Developing **Spark scripts, UDF's** using both **Spark DSL** and Spark SQL query for data aggregation, querying, and writing data back into RDBMS through **Sqoop.**
  + Used Snowflake to separate compute usage from storage in their pricing structure.
* Developed **Spark** code using**Scala** and **Spark-SQL/**Streaming for faster processing of data.
* Create and maintain the data pipelines using matilion ETL, Fivetran
* Evaluated Fivetran and matilion for streaming and batch data ingestion into snowflake
  + Ingested data into HDFS using **SQOOP** and scheduled an incremental load to **HDFS.**
  + Collaborated with business analysts to understand data requirements and translate them into effective PL SQL solution
* Using Hive to analyze data ingested into **HBase** by using **Hive-HBase** integration and compute various metrics for reporting on the dashboard.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory Palantir Foundry, Spark SQL and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services- (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in InAzure Data bricks.
* Developed a Spark job in Java which indexes data into Elasticsearch from external Hive tables which are in HOFS
  + Utilized Apache Spark with **Python** to develop and execute Big Data Analytics and Machine learning applications, executed machine Learning use cases under **Spark** ML and Mllib
  + AWS CI/CD Data pipeline and AWS Data Lake using EC2, AWS Glue, AWS Lambda.
  + Code review of all the Kafka Unit test case documents, Palantir Foundry, Talend, EQA documents, completed by team with proper review check list. Also do development for same.
  + Used to store relational database and works well with Tableau, Excel and many other tools familiar to end users.
  + Data sources are extracted, transformed and loaded to generate CSV data files with **Python** programming and SQL queries.
* Worked on **NiFi** data Pipeline to process large set of data and configured Lookup’s for Data Validation and Integrity.
  + Worked with Hadoop infrastructure to storage data in **HDFS storage** and use **Spark / HIVE SQL** to migrate underlying SQL codebase in **Azure.**
  + Maintained and debugged existing PL SQL codebase, resolving issues promptly.
  + Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS, HBase

and ElasticSearch.

* + Wrote Pig Scripts to generate **Map Reduce** jobs and performed ETL procedures on the data in **HDFS.**
  + Data Stack: Azure Databricks, ADLS, ADF, AAS, DAX, Azure Automation Accounts, Azure Active Directory(AD), Azure IAM security Groups, Pyspark, Spark SQL, Azure Data warehouse (ADW), Power BI, DAX coding, MSBI, SSAS, CI/CD and Production Support.
  + **Knime tools of visualization widely used :Microstrategy< tableau,powerbi,waston analytics,dbvisulizer**
  + Experienced in loading the real-time data to **NoSQL** database like **Cassandra.**
  + Developing scripts in Pig for transforming data and extensively used event joins, filtered and done pre- aggregations.
  + Performed Data scrubbing and processing with**Apache Nifi**and for workflow automation and coordination.
  + Used Sqoop to import data into **HDFS** and **Hive** from Oracle database.
  + Developed code in Java which creates mapping in ElasticSearch even before data is indexed into.
  + Used Snowflake for fundamentally built to be a complete SQL database.
  + Worked on implementation of a log producer in Scala that watches for application logs, transform incremental log and sends them to a Kafka and Zookeeper based log collection platform.
  + Using apache **NiFi** to copy the data from local file system to HDFS.
* Used **Talend**for Big data Integration using **Spark**andHadoop.
* Worked in developing Pig Scripts for data capture change and delta record processing between newly arrived data and already existing data in HDFS.
* Performed end-to-end delivery of pyspark ETL pipelines on Azure-databricks to perform the transformation of data orchestrated via Azure Data Factory (ADF) scheduled through Azure automation accounts and trigger them using Tidal Schedular.
  + Optimized Hive queries to extract the customer information from **HDFS.**
  + Involved in various phases of development analyzed and developed the system going through **Agile Scrum methodology.**
* Generate **metadata,** create Talend etl jobs**,** mappings to load data warehouse**, data lake.**
* Used **Zookeeper** to provide coordination services to the cluster.
  + Analyzed data using **Hive** the partitioned and bucketed data and compute various metrics for reporting.
  + Built **Azure** Data Warehouse Table Data sets for **PowerBI Reports.**
  + Import data from sources like **HDFS/HBase** into **Spark RDD.**
  + Good experience in developing Hive **DDLs** to create, alter and drop Hive TABLES.
* Working on **BI reporting** with At Scale **OLAP** for **Big Data**.
  + Implement enterprise grade platform(mark logic) for ETL from mainframe to **NOSQL**(Cassandra).
  + Implemented **Kafka** for streaming data and filtered, processed the data.
* Designed and Developed Real time Stream processing Application using **Spark, Kafka, Scala** and **Hive** to perform Streaming ETL and apply Machine Learning.
* Developed complex PL SQL scripts and queries for data manipulation and analysis.
  + Developed data pipeline using flume, **Sqoop** and pig to extract the data from weblogs and store in **HDFS.**
  + Developed Shell scripts for scheduling and automating the job flow.
  + Developed a workflow using **Nifi** to automate the tasks of loading the data into HDFS.
  + Developed **MapReduce** jobs to calculate the total usage of data by commercial routers in different locations, developed Map reduce programs for data sorting in HDFS
  + Load balancing of ETL processes, database performance tuning ETL processing tools.
  + Loaded the data from **Teradata** to **HDFS** using Teradata Hadoop connectors.

**Environment:** Spark, YARN, HIVE, Pig, Scala, Mahout, NiFi, Python, Hadoop, Azure, Dynamo DB, Kibana, NOSQL, Sqoop, MYSQL.

**Client: TFS, Plano, TX Aug 2016 to Sep 2018 Role: Sr. Big Data Engineer**

**Responsibilities:**

* Installed **Hadoop, Map Reduce, HDFS, AWS** and developed multiple **Map Reduce** jobs in **PIG** and **Hive** for **data cleaning** and **pre-processing.**
* Implemented **Spark** using **Scala** and utilizing Data frames and **Spark SQL API,** Data Frames and **Pair RDD's** for faster processing of data and created RDD's, Data Frames and datasets.
* Worked extensively on building Nifi data pipelines in docker container environment in development phase.
* Developed **PIG** scripts to transform the raw data into intelligent data as specified by business users.
  + Designed both **3NF** data models for **ODS, OLTP, OLAP** systems and dimensional data models using star and snow flake Schemas.
  + Utilized Palantir Foundryand Docker for the runtime environment of the CI/CDsystem to build, test deploy. Analysed, design and build Modern data solutions using Azure PaaS service to support visualization of data. Understand current Production state of application and determine the impact of new implementation on existing business processes.
  + Implemented Installation and configuration of multi-node cluster on Cloud using Amazon Web Services (AWS) on **EC2.**
  + Extensive experience in writing **UNIX shell scripts** and automation of the ETL processes using UNIX shell scripting.
* Implemented solutions for ingesting data from various sources and processing the Data-at-Rest utilizingBig Data technologies such as**Hadoop, Map Reduce Frameworks, HBase, Hive .**
* Installed and configured Hadoop MapReduce, HDFS, developed multiple MapReduce jobs in Java and Nifi for data cleaning and preprocessing.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
  + Involved in scheduling **Oozie** workflow engine to run multiple **Hive** and **Pig jobs.**
  + Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using **Spark Context**, Spark-SQL, **Data Frame, Pair RDD's, Spark YARN.**
  + Worked on batch processing of data sources using **Apache Spark**, Elastic search.
  + Involved in converting **Hive/SQL queries** into Spark transformations using **Spark RDDs, Scala.**
  + Worked on migrating **PIG scripts** and MapReduce programs to Spark Data frames API and Spark SQL to improve performance.
  + Expertise in implementing DevOps culture through CI/CD tools like Repos, Code Deploy, Code Pipeline, GitHub.
  + Ingest data into **Hadoop / Hive/HDFS** from different data sources.
  + Created Hive External tables to stage data and then move the data from Staging to main tables.
  + Created Data Pipelines as per the business requirements and scheduled it using **Oozie** Coordinators.
  + Worked with **NoSQL** database **HBase** in getting real time data analytics.
  + Able to assess business rules, collaborate with stakeholders and perform source-to-target data mapping, design and review.
  + Used Oozie workflow engine to manage interdependent Hadoop jobs and to automate several types of Hadoop jobs such as **MapReduce Hive, Pig**, and **Sqoop.**
  + Finalize the naming Standards for Data Elements and ETL Jobs and create a Data Dictionary for Meta Data Management.
  + Created scripts for importing data into **HDFS/Hive** using **Sqoop** from **DB2.**
  + Loading data from different source(database & files) into Hive using **Talend tool.**
  + Involved in collecting, aggregating and moving data from servers to HDFS using Apache Flume.
  + Objective of this project is to build a data lake as a cloud based solution in **AWS** using **Apache Spark.**
  + Developed **Pig scripts** to parse the raw data, populate staging tables and store the refined data in partitioned DB2 tables for Business analysis.
  + Worked on managing and reviewing Hadoop log files. Tested and reported defects in an Agile Methodology perspective.
  + Conduct/Participate in project team meetings to gather status, discuss issues & action items
  + Provide support for research and resolution of testing issues.
  + Coordinating with Business for UAT sign off.
  + Designed, developed and maintained data integration programs in Hadoop and RDBMS environment with both RDBMS and **NoSQL** data stores for data access and analysis.
  + Used all major ETL transformations to load the tables through Informatica mappings.
  + Created **Hive queries** and tables that helped line of business identify trends by applying strategies on historical data before promoting them to production.
  + Worked on Data modeling, **Advanced SQL** with Columnar Databases using **AWS.**

**Environment:** Hadoop, Cloudera, Talend, Scala, Spark, HDFS, Hive, Pig, Sqoop, DB2, SQL, Linux, Yarn, NDM, Informatica, AWS, Windows & Microsoft Office.Fivetran

**Client: Dow Jones, Monmouth Junction, NJ Sep 2015 to July 2016 Role: Sr. Data Modeler/ Analyst**

**Responsibilities:**

* Data Modeler/Analyst in Data Architecture Team and responsible for Conceptual, Logical and Physical model for Supply Chain Project.
* Created and maintained **Logical** &**Physical Data Models** for the project. Included documentation of all entities, attributes, data relationships, primary and foreign key structures, allowed values, codes, business rules, glossary terms, etc.
* Created conceptual, logical and physical data models, data dictionaries, **DDL** and **DML** to deploy and load database table structures in support of system requirements.
  + Generated ad-hoc **SQL queries** using joins, database connections and transformation rules to fetch data from legacy DB2 and **SQL Server** database systems.
  + Translated business requirements into working **logical** and **physical data models** for **OLTP**&**OLAP** systems.
  + Worked on normalization techniques. Normalized the data into 3rd Normal Form **(3NF).**
  + Creation of **BTEQ, Fast export, Multi Load, TPump, Fast load** scripts for extracting data from various production systems.
* Owned and managed all changes to the data models. Created data models, solution designs and data architecture documentation for complex information systems.
* Developed Advance **PL/SQL packages, procedures, triggers, functions**, Indexes and Collections to implement business logic using **SQL Navigator.**
  + Involved in designing data warehouses and data lakes on regular **(Oracle, SQL Server)** high performance on big data (**Hadoop - Hive** and **HBase**) databases. Data modeling, Design, implement, and deploy high-performance, custom applications at scale on **Hadoop /Spark.**
  + Involved in Extract, Transform and Load (ETL) data from spreadsheets, flat files, database tables and other sources using SQL Server Integration Services **(SSIS)** and SQL Server Reporting Service **(SSRS)** for managers and executives.
  + Designed Star Schema Data Models for **Enterprise Data Warehouse** using Power Designer.
  + Created Mapping documents for Staging, ODS &**Data Mart Layers.**
  + Created **PL/SQL packages** and Database Triggers and developed user procedures and prepared user manuals for the new programs.
  + Worked on the **OLAP** for data warehouse and data mart developments using Ralph Kimball methodology as well as **OLTP models**, both and interacting with all the involved stakeholders and SME's to derive the solution.
  + Created Model reports including Data Dictionary, Business reports.
  + Generated **sql scripts** and implemented the relevant databases with related properties from keys, constraints, indexes & sequences.
  + Performed gap analysis and dependency analysis for current & future systems.
* Created Communities, Domains, Assets, hierarchies in **Collibra.**
  + Reviewed **Stored Procedures** for reports and wrote test queries against the source system (**SQL Server-SSRS**) to match the results with the actual report against the **Datamart** (Oracle).
* Spearheaded the establishment of the Enterprise Business Glossary, including Business Terms, BT Descriptions, and Business Rules; the Tiering Criteria, encompassing Tier 1, 2 or 3; and the Data Linkages between the Metadata and Lineage documents for **Collibra, IDQ**, and **IMM data governance tools**. Integrated process to manage data quality
* Preparation of business **(Collibra)** and technical **metadata (IBM Infosphere)**
  + Created Use Case Diagrams using **UML** to define the functional requirements of the application.
  + Created the best fit Physical Data Model based on discussions with DBAs and ETL developers.
  + Identified required dimensions and Facts using **Erwin tool** for the Dimensional Model.
  + Implemented ETL techniques for Data Conversion, **Data Extraction** and **Data Mapping** for different processes as well as applications.
  + Created the best fit Physical Data Model based on discussions with DBAs and ETL developers.
  + Designed ER diagrams (**Physical and Logical using Erwin**) and mapping the data into database objects.
* Validated and updated the appropriate Models to process mappings, screen designs, use cases, business object model, and system object model as they evolved and changed.

**Environment:** OLTP, DBAs, DDL, DML, Erwin, UML, diagrams, Snow-flak schema, SQL, Data Mapping, Metadata, OLTP, SAS, Informatica 9.5,AWS

**Client: Keystone Logic, IND Feb2012 to Aug 2015 Role: Big Data Developer**

**Responsibilities:**

* As a Sr. Data Modeler/Data Analyst I am responsible for all data related aspects of a project.
* Worked on Software Development Life Cycle **(SDLC)** with good working knowledge of testing**, Agile methodology**, disciplines, tasks, resources and scheduling.
* Developed normalized Logical and Physical database models to design **OLTP** system for Reference and Balance data conformance using **ER studio modeling tool**.
* Worked with SQL, Python, Oracle PL/SQL **, Stored Procedures, Triggers, SQL queries** and loading data into Data Warehouse/Data Marts.
* Developed the logical data models and physical data models that capture current state/future state data elements and data flows using **ER Studio**.
* Involved in preparing the design flow for the **Data stage objects** to pull the data from various upstream applications and do the required transformations and load the data into various downstream applications.
* Worked in importing and cleansing of data from various sources like **Teradata, Oracle, flat files, SQL Server** with high volume data.
* Delivered dimensional data models using **ER/Studio** to bring in the Employee and Facilities domain data into the oracle data warehouse.
* Performed analysis of the existing source systems (Transaction database)
* Involved in maintaining and updating Metadata Repository with details on the nature and use of applications/data transformations to facilitate impact analysis.
* Created DDL scripts using **ER Studio** and source to target mappings to bring the data from source to the warehouse.
* Business Data Lineage from Critical Data Elements to DQ Measures to Business Rules mapped on **Collibra**
* Designed the ER diagrams, logical model (relationship, cardinality, attributes, and, candidate keys) and physical database (capacity planning, object creation and aggregation strategies) for **Oracle** and **Teradata**.
* Worked in importing and cleansing of data from various sources like **Teradata**, **Oracle**, **flatfiles**, **MS SQL Server** with high volume data
* Reverse Engineered DB2 databases and then forward engineered them to Teradata using **ER Studio.**
* Part of team conducting logical data analysis and **data modeling JAD sessions**, communicated data-related standards .
* Involved in meetings with SME (subject matter experts) for analyzing the multiple sources.
* Involved in SQL queries and optimizing the queries in **Teradata.**
* Created **DDL scripts using ER Studio** and source to target mappings to bring the data from source to the warehouse.
* Wrote and executed SQL queries to verify that data has been moved from transactional system to **DSS, Data warehouse, data mart** reporting system in accordance with requirements.
* Worked extensively on **ER Studio** for multiple Operations across Atlas Copco in both **OLAP** and **OLTP** applications.
* Used forward engineering to create a physical data model with **DDL** that best suits the requirements from the **Logical Data Model**.
* Worked with the DBA to convert logical Data models to physical Data models for implementation.

**Environment:** Business Objects, ER Studio, Oracle SQL Developer,AWS, SQL Server 2008, Teradata, ER/Studio, SSIS, Windows, MS Excel.

**Environment**: Erwin 4, MS Visio, Oracle 10g, SQL Server 2000, Business Object Data Integrator R2

**Education Details:-** Completed Bachelor’s in 2011 from India