NAME: Keerthi Sai Sri
EMAIL: keerthisaisri14@gmail.com
CONTACT: 913-662-2316

Senior Data Engineer

BACKGROUND SUMMARY:

- 10 years of experience in data analysis, data engineering, and statistical modeling, including data extraction, manipulation, visualization, and validation techniques, and reporting on various projects.
- Worked on Scala codebase related to Apache Spark performing the Actions, Transformations on RDDs, Data Frames & Datasets using Spark SQL and Spark Streaming Contexts.
- Experience in data architecture including data ingestion pipeline design, Hadoop information architecture, data modeling, data mining, machine learning, and advanced data processing.
- Expertise in creating Pods using Kubernetes and worked with Jenkins pipelines to drive all microservices builds out to the Docker registry and then deployed to the Kubernetes cluster.
- A passionate data professional with experience in designing and implementing Data Warehousing concepts, Business Intelligence projects for Health care, and Banking and ETL solutions.
- Good usage of Apache Hadoop along with the enterprise version of Cloudera and Hortonworks.
- Hands on experience with Big Data Ecosystems including Hadoop, MapReduce, Pig, Hive, Impala, Sqoop, Flume, NIFI, Oozie, MongoDB, Zookeeper, Kafka, Maven, Spark, Scala, HBase, Cassandra.
- Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on-premises ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud DataProc, Google Cloud Storage, Composer.
- Experience in building data pipelines using Azure Data factory, Azure data bricks and loading data to Azure Data Lake, Azure SQL Database, Azure SQL Data warehouse and controlling and granting database access.
- Strong experience working with Amazon cloud services like EMR, Redshift, DynamoDB, Lambda, Athena, Glue, S3, API Gateway, RDS, CloudWatch for efficient processing of Big Data.
- Experience on Migrating SQL database to Azure Data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and controlling and granting database access and Migrating On premise databases to Azure Data Lake store using Azure Data factory.
- Good experience on Snowflake. Connected to the AWS Redshift and Snowflake directly from Tableau.
- Experienced in integrating Kafka with Spark Streaming for real-time data processing.
- Experience in Database Design and development with Business Intelligence using SQL Server 2014/2016, Integration Services (SSIS), DTS Packages, SQL Server Analysis Services (SSAS), DAX, OLAP Cubes, Star Schema and Snowflake Schema.
- Experience on Physical & Logical Data Modelling, Dimensional Modelling using Star and Snowflake Schemas, Data marts, OLAP, FACT & Dimensions tables.
- Extensively worked with Teradata utilities Fast export, and Multi Load to export and load data to/from different source systems including flat files.
- Experienced in ingesting data into HDFS from various Relational databases like MYSQL, Oracle, DB2, Teradata, Postgres using Sqoop.
- Experience in Python, Scala, shell scripting, and Spark.
- Proficient in data processing like collecting, aggregating, moving from various sources using Apache Flume and Kafka.
- 4+ years of Strong expertise in using ETL Tool Informatica Power Center 10.x/9.x/8.x (Designer, Workflow Manager, Repository Manager, ETL and Data Warehouse.
- Experience in Developing Spark applications using Spark - SQL, Pyspark and Delta Lake in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Experience with MS SQL Server Integration Services (SSIS), T-SQL skills, stored procedures, triggers.

- Extensive experience in developing stored procedures, functions, Views and Triggers, Complex queries using SQL Server, TSQL and Oracle PL/SQL

TECHNICAL SKILLS:

Big Data Technologies	Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Zookeeper, Apache Flume, Apache Airflow, Cloudera, HBase
Programming Languages	Python, PL/SQL, SQL, Scala, C, C#, C++, T-SQL, Power Shell Scripting, JavaScript
Cloud Services	Azure Data Lake Storage Gen 2, Azure Data Factory, Blob storage, Azure SQL DB, Databricks, Azure Event Hubs, AWS RDS, Amazon SQS, Amazon S3, AWS EMR, Lambda, AWS SNS.
Databases	MySQL, SQL Server, Oracle, MS Access, Teradata, and Snowflake
NoSQL Data Bases	MongoDB, Cassandra DB, HBase
Development Strategies	Agile, Lean Agile, Pair Programming, Waterfall and Test Driven Development.
Visualization & ETL tools	Tableau, Informatica, Talend, SSIS, and SSRS
Version Control & Containerization tools	Jenkins, Git, and SVN
Operating Systems	Unix, Linux, Windows, Mac OS
Monitoring tool	Apache Airflow, Control M

PROFESSIONAL EXPERIENCE:

Role: Big Data Engineer
Client: Citi Bank, Charlotte, NC
Duration: March 2021 – Present
Responsibilities:

- Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats.
- Worked on building the data pipelines (ELT/ETL Scripts), extracting the data from different sources (MySQL, AWS S3 files), transforming, and loading the data to the Data Warehouse (AWS Redshift)
- Extensive experience in working with AWS cloud Platform (EC2, S3, EMR, Redshift, Lambda and Glue).
- Worked on developing & adding few Analytical dashboards using Looker product.
- Worked on building the data pipelines using PySpark (AWS EMR), processing the data files present in S3 and loading it to Redshift.
- Successfully completed a POC on GCP services such as Big Query, Dataflow, Pub/Sub, and Cloud Storage, demonstrating the ability to quickly learn and work with new cloud platforms.
- Used Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra
- Created ETL Mapping with Talend Integration Suite to pull data from Source, apply transformations, and load data into target database.

- Proficient in working with Databricks notebooks to develop, test, and deploy ETL pipelines using languages such as Python, Scala, or SQL. Developed ETL's using PySpark. Used both Data frame API and Spark SQL API
- Experience designing, building, and maintaining Hadoop-based data processing systems using Scala, including expertise in Spark and related technologies such as HDFS, Hive, and Kafka.
- Strong understanding of distributed computing principles and experience working with large datasets, as well as experience with data modeling, schema design, and database management.
- Experience with AWS services related to data processing, such as Amazon EMR, Amazon S3, and AWS Glue, as well as proficiency in scripting languages such as Python and Bash for data processing and automation tasks.
- Experience in data profiling, data mapping, data cleaning, data integration, meta data management and MDM (Master Data Management)
- Expertise in integrating Databricks with other tools such as Apache Spark, Apache Hadoop, and cloud services like AWS.
- Authoring Python (PySpark) Scripts for custom UDF's for Row/ Column manipulations, merges, aggregations, stacking, data labeling and for all Cleaning and conforming tasks.
- Stored the log files in AWS S3. Used versioning in S3 buckets where the highly sensitive information is stored.
- Integrated AWS Dynamo DB using AWS lambda to store the values of items and backup the DynamoDB streams.
- Prepared scripts to automate the Ingestion process using Pyspark and Scala as needed through various sources such as API, AWS S3, Teradata and Redshift.
- Involved in designing different components of system like Sqoop, Hadoop process involves map reduce & hive, Spark, FTP integration to down systems.
- Using Spark, performed various transformations and actions and the result data is saved back to HDFS from there to target database Snowflake
- Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift
- Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
- Worked on scheduling all jobs using Airflow scripts using python added different tasks to DAG, LAMBDA.
- Used Pyspark for extract, filtering and transforming the Data in data pipelines.
- Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.

Environment: Python, Spark, AWS EC2, AWS S3, AWS EMR, AWS Redshift, AWS Glue, AWS RDS, AWS Kinesis firehose, kinesis data stream, AWS SNS, AWS SQS, AWS Athena, snowflake, SQL, Tableau, Git, REST, Bitbucket, Jira.

Role: Azure Data Engineer
Client: T Mobile, Overland Park, Kansas
Duration: Dec 2018 – Feb 2021

Responsibilities:

- Develop, and design data models, data structures, and ETL jobs for data acquisition and manipulation purposes.
- Architect & implement medium to large-scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB, Cosmos DB).
- Create pipelines in ADF using linked services to extract, transform and load data from multiple sources like Azure SQL, Blob storage and Azure SQL Data warehouse.
- Develop batch processing solutions by using Data Factory and Azure Databricks.
- Performed ETL operations in Azure Databricks by connecting to different relational database source systems using JDBC connectors.

- Developed Python scripts to do file validations in Databricks and automated the process using ADF.
- Developed an automated process in Azure cloud that can ingest data daily from web service and load it into Azure SQL DB.
- Used Spring Kafka API calls to process the messages smoothly on Kafka Cluster setup.
- Involved in all the steps and scope of the project reference data approach to MDM, have created a Data Dictionary and Mapping from Sources to the Target in MDM Data Model.
- Create pipelines in ADF using linked services to extract, transform and load data from multiple sources like Azure SQL, Blob storage and Azure SQL Data warehouse.
- Developed Streaming pipelines using Azure Event Hubs and Stream Analytics to analyze data for dealer efficiency and open table counts for data coming in Fiot-enabled poker and other pit tables.
- Developed custom alerts using Azure Data Factory, SQLDB, and Logic App.
- Developed Databricks ETL pipelines using notebooks, Spark Data frames, SPARK SQL, and Python scripting. Design for data auditing and data masking
- Developed multi-cloud strategies in better using GCP (for its PASS) and Azure (for its SAAS).
- Experience in moving data between GCP and Azure using Azure Data Factory.
- Performed Application-level DBA activities creating tables, and indexes, and monitored and tuned Teradata BETQ scripts using Teradata Visual Explain utility.
- Integrated both framework and CloudFormation to automate Azure environment creation along with the ability to deploy on Azure, using build scripts (Azure CLI) and automate solutions using Terraform.
- Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, 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 In Azure Databricks.
- Performance tuning, monitoring, UNIX shell scripting, and physical and logical database design.
- Performed ETL operation using SSIS and loaded the data into Secure DB.
- Good hands-on experience in Data Vault concepts, and data models, well-versed understanding, and implementation of Data warehousing concepts/Data Vault.
- Designed, reviewed, and created primary objects such as views, and indexes based on logical design models, user requirements, and physical constraints.
- Worked with stored procedures for data set results for use in Reporting Services to reduce report complexity and optimize the run time. Exported reports into various formats (PDF, Excel) and resolved formatting issues.

Environment: Azure Data Factory, Shell Scripting, Teradata, python scripting, Azure data bricks, Azure data lake storage, Blob storage, Azure SQL Database, snowflake, Azure Synapse analytics, Azure Synapse workspace, Synapse SQL pool, Power BI, HDInsight.

Role: Hadoop Developer
Client: Bed Bath & Beyond, Union, NJ
Duration: Aug 2016 – Nov 2018

Responsibilities:

- Proactively monitored systems and services, architecture design and implementation of Hadoop deployment, configuration management, backup, and disaster recovery systems and procedures and
- Worked on analyzing Hadoop cluster using different big data analytic tools including Kafka, Pig, Hive and MapReduce.
- Configured Spark streaming to receive real time data from the Kafka and store the stream data to HDFS using Scale.
- Installed and configured Hadoop, MapReduce, HDFS (Hadoop Distributed File System), developed multiple MapReduce jobs in java for data cleaning and processing.
- Worked on implementing Spark using Scala and SparkSQL for faster analyzing and processing of data.
- Used JAVA, J2EE application development skills with Object Oriented Analysis and extensively involved throughout Software Development Life Cycle (SDLC)

- Implemented AWS EC2, Key Pairs, Security Groups, Auto Scaling, ELB, SQS, and SNS using AWS API and exposed as the Restful Web services.
- Involved in creating Hive tables, loading the data and writing hive queries, which will run internally in map reduce and applied MapReduce framework jobs in java for data processing by installing and configuring Hadoop, HDFS.
- Responsible for developing data pipeline using flume, Sqoop and pig to extract the data from weblogs and store in HDFS and involved in developing Pig Scripts for change data capture and delta record processing between newly arrived data and already existing data in HDFS.
- Implemented Reporting, Notification services using AWS API and used AWS (Amazon Web services) compute servers extensively.
- Written Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data and involved in scheduling Oozie workflow engine to run multiple Hive and pig jobs.
- Worked on Designing and Developing ETL Workflows using Java for processing data in HDFS/Hbase using Oozie.
- Create Snapshots of EBS Volumes and Monitor AWS EC2 Instances using Cloud Watch and worked on AWS Security Groups and their rules
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala and generated JavaAPIs for retrieval and analysis on No-SQL database such as HBase and worked with NoSQL databases like HBase in creating tables to load large sets of semi structured data.
- Worked on loading data from UNIX file system to HDFS and analyzed large amounts of data sets to determine optimal way to aggregate and report on it.
- Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team Using Tableau.

Environment: Hadoop, Java/J2EE, HDFS, MapReduce, AWS, EC2, RDS, S3, Cloud Watch, Hive Sqoop, Pig, HBase, Apache Spark, Oozie Scheduler, Java, UNIX Shell Scripts, Kafka, Git, Maven, PLSQL, MongoDB, HBase, Cassandra, Python, Scala, Teradata, Netezza, Oracle.

Role: Hadoop Developer
Client: Global Edge Software, India
Duration: Aug 2015 – May 2016

Responsibilities:

- Actively Participated in all phases of the Software Development Life Cycle (SDLC) from implementation to deployment.
- Responsible for building scalable distributed data solutions using Hadoop.
- Responsible for Cluster maintenance, adding and removing cluster nodes, Cluster Monitoring and Troubleshooting, Managing, and reviewing data backups & log files.
- Responsible to manage the test data coming from different sources.
- Analyzed data using Hadoop components Hive and Pig.
- Load and transform large sets of structured, semi structured, and unstructured data using Hadoop/Big Data concepts.
- Involved in importing and exporting the data from RDBMS to HDFS and vice versa using Sqoop.
- Involved in loading data from UNIX file system to HDFS.
- Responsible for creating Hive tables, loading data, and writing hive queries.
- Created Hive External tables and loaded the data into tables and query data using HQL.
- Handled importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS.
- Created and maintained technical documentation for launching Hadoop Clusters and for executing Hive queries and Pig Scripts.
- Extracted the data from Teradata into HDFS using the Sqoop.
- Exported the patterns analyzed back to Teradata using Sqoop.

- Experience in Monitoring System Metrics and logs for any problems adding, removing, or updating Hadoop Cluster.
- Involved in scheduling Oozie workflow engine to run multiple Hives and pig jobs and used Oozie workflows for batch processing and scheduling workflows dynamically.
- Involved in requirement analysis, design, coding, and implementation phases of the project.

Environment: Hadoop, Spark, Scala, MapReduce, HDFS, Hive, Java, SQL, Cloudera Manager, Pig, Sqoop, Oozie, Zookeeper

Role: SQL Server Developer
Client: Enterpi Software solutions, India
Duration: Apr 2014 – July 2015

Responsibilities:

- Coordinated with front-end application developers for implementing database architecture and design.
- Used various transformations in SSIS dataflow, control flow using for loop containers, and fuzzy lookups.
- Develop parameterized reports, caching reports, sub reports and Ad Hoc reports using SSRS.
- Implemented error handling and utilized event handlers for automated notifications using SSIS.
- Write Complex SQL Queries to generate Reports based on the business requirement.
- Redesigned the SSIS packages from the legacy DTS packages.
- Execute SSIS package include a master package which include number of child packages.
- Supporting ETL (Extract Transform and Load) for fetching data from multiple systems to single Data Warehouse.
- Created complex Ad-Hoc reports, Sub reports, linked reports related to State compliance reporting. Used custom code in SSRS for row color, visibility, and masking.
- Developed Full Analysis Cycle Project and created packages for extracting data from OLTP to OLAP. Created Multi-Dimensional Expression (MDX) scripts for OLAP data cubes.
- Involved in Designing, Developing and Testing of the ETL (Extract, Transformation and Load) strategy to populate the data from Heterogeneous data sources (SQL Server, Flat Files, Excel source files, XML files etc.)
- Performed different kinds of transformations like Lookup Transformations, Merge Joins, Derived Columns, Merge Join, Conditional Split, and Data Conversion with Multiple Data Flow tasks.
- Bulk data migration using Bulk Insert from flat files.
- Created Package Configurations, Event Handlers for On Error, On Pre/Post execution. Designed Complex Packages with Error Handling and Package Logging that stores the Logging results in SQL Tables and log files Performance tuning and testing on stored procedures, indexes using performance tool like SQL Profiler and DETA.
- Import/Export and successfully migrated the Server usage MS Access database data to SQL Server using ETL/SSIS. Created dynamic and customized packages to support future changes. Scheduled the same packages by creating the corresponding job tasks.
- Formatted the reports using the global variables and expressions in SSRS, deployed the generated reports on to the Reporting server.

Environment: MS SQL Server 2012, Visual Studio 2010, T-SQL, MS Excel, Microsoft SQL Server Integration Services (SSIS), Microsoft SQL Server Reporting Services (SSRS), Rally bug tracking and SVN.

EDUCATION:

Bachelor's - Andhra Loyola Institute of Engineering and Technology
Major: Computer Science
Aug 2010 - March 2014