Silvister. T

Phone: +1 8329150573

Email: silvistert7@gmail.com

**Professional Summary:**

Over 8+ years of strong experience in Data Analyst, Data mining with large data sets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive modeling, Statistical modeling, Data modeling, Data Visualization. Adept in statistical programming languages like R and Python, SAS, Apache Spark including Big Data technologies like Hadoop, Hive, Pig.

* 3+ years of experience in Hadoop 2.0. Led development of enterprise level solutions utilizing Hadoop utilities such as Spark, MapReduce, Sqoop, PIG, Hive, HBase, Zookeeper, Phoenix, Oozie, Flume, streaming jars, Custom Ser De, etc. Worked on proof of concepts on Kafka, and Storm.
* Deep analytics and understanding of Big Data and algorithms using Hadoop, MapReduce, NoSQL and distributed computing tools.
* Experience in developing enterprise level solution using batch processing (using Apache Pig) and streaming framework (using Spark Streaming, apache Kafka & Apache Flink).
* Expertise in synthesizing Machine learning, Predictive Analytics and Big data technologies into integrated solutions.
* Experienced in Dimensional Data Modeling experience using Data modeling, Relational Data modeling, ER/ Studio, Erwin, and Sybase Power Designer, Star Join Schema/Snowflake modeling, FACT & Dimensions tables, Conceptual, Physical & logical data modeling.
* Procedural knowledge in cleansing and analyzing data using HiveQL, Pig Latin, and custom MapReduce programs in Java.
* Designed and developed ETL packages using SQL Server Integration Services (SSIS) to load the data from SQL server, XML files to SQL Server database through custom C# script tasks.
* Experienced in writing custom UDFs and UDAFs for extending Hive and Pig core functionalities.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems (RDBMS), Teradata and vice versa.
* Strong experience in design and development of Business Intelligence solutions using data modeling, Dimension Modeling, ETL Processes, Data Integration, OLAP and client /server application.
* Worked extensively with Dimensional modeling, Data migration, Data cleansing, Data profiling, and ETL Processes features for data warehouses.
* Hands on experience in formatting and ETL raw data in various format such as Avro, ORC, Parquet, CSV, JSON etc. Experience in Elasticsearch and MDM solutions.
* Excellent knowledge of Machine Learning, Mathematical Modeling and Operations Research. Comfortable with R, Python, SAS and, Relational databases. Deep understanding & exposure of Big Data Ecosystem.
* Extensively used SQL, NumPy, Pandas, Sci kit - learn, Spark, Hive for Data Analysis and Model building.
* Have experience with processing Framework such as Spark and Spark SQL.
* Experienced on data architecture including data ingestion pipeline design, Hadoop information architecture, data modeling and data mining, machine learning and advanced data processing.
* Analysis and Factor Analysis, testing and validation using ROC plot, K- fold cross validation and data visualization.
* Extensively used SQL, NumPy, Pandas, Sci kit-learn, Spark, Hive for Data Analysis and Model building.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Experience in building efficient pipelines for moving data between GCP and Azure using Azure Data Factory.
* Experience in building power bi reports on Azure Analysis services for better performance when comparing that to direct query using GCP Big Query.
* Hands on experience in implementing LDA, Naive Bayes and skilled in Random Forests, Decision Trees, Linear and Logistic Regression, SVM, Clustering, neural networks, Principal Component Analysis.
* Solid experience on building ETL ingestion flows using Azure Data Factory.
* Experience in building the data pipeline by leveraging the Azure Data Factory.
* Hands on experience working with SSIS, for ETL process ensuring proper implementation of Event Handlers, Loggings, Checkpoints, Transactions and package configurations.
* Hands - on experience on Google Cloud Platform (GCP) in all the big data products big query, Cloud Data Pro, Google Cloud Storage, Composer (Air Flow as a service).
* Expertise in managing entire data science project life cycle and actively involved in all the phases of project life cycle including data acquisition, data cleaning, data engineering, features scaling, features engineering, statistical modeling (decision trees, regression models, neural networks, SVM, clustering), dimensionality reduction using Principal Component Analysis.
* Strong experience and knowledge in Data Visualization with Tableau creating: Line and scatter plots, Bar Charts, Histograms, Pie chart, Dot charts, Box plots, Time series, Error Bars, Multiple Charts types, Multiple Axes, subplots etc.
* Experienced with Integration Services (SSIS), Reporting Service (SSRS) and Analysis Services (SSAS)
* Expertise in Normalization to 3NF/De-normalization techniques for optimum performance in relational and dimensional database environments.
* Well-versed in version control and CI-CD tools such as GIT, SourceTree, Bitbucket, etc., and Amazon Web Services (AWS) products S3, EC2, EMR, and RDS.
* Experience in all stages of SDLC (Agile, Waterfall), writing Technical Design document, Development, Testing and Implementation of Enterprise level Data mart and Data warehouses.

**Technical Skills:**

|  |  |
| --- | --- |
| OLAP/Reporting Tools | SSRS, SSAS, MDX, Tableau, Power BI |
| Relational Databases | SQL Server database 2014/2012/2008 R2/2005, Oracle 11g, SQL Server Azure, MS Access |
| SQL Server Tools | Microsoft Visual Studio 2010/2013/2015, SQL server management studio |
| Big Data Ecosystem | HDFS, NIFI, Map Reduce, Oozie, Hive/Impala, Pig, Sqoop, Zookeeper and H base, Spark, Scala, Kafka, Apache Flink, AWS- EC2, S3, EMR. |
| Other Tools | MS Office 2003/2007/2010 and 2013, Power pivot, Power Builder, GIT, CI-CD, Jupiter Note Book. |
| Google Cloud Platform | GCP Cloud Storage, Big Query, Azure, Cloud Data, SQL, Cloud Functions, Pub/sub |
| Programming languages | C, SQL, PL/SQL, T-SQL, JAVA, Batch scripting, R, Python |
| Data Warehousing & BI | Star Schema, Snowflake schema, Facts and Dimensions tables, SAS, SSIS, and Splunk |
| Operating Systems | Windows XP/Vista/7/8 and 10; Windows 2003/2008R2/2012 Servers |

**Professional Experience:**

Client: Health Catalyst, Utah August 2022 - Present

Role: Data Engineer

**Responsibilities:**

* Involved in requirements gathering, analysis, design, development, change management, deployment.
* Involved in the development of real time streaming applications using PY Spark, Apache Flink, Kafka, Hive on distributed Hadoop Cluster.
* 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 ML lib.
* Extracted data from heterogeneous sources and performed complex business logic on network data to normalize raw data which can be utilized by BI teams to detect anomalies.
* Designed and developed Flink pipelines to consume streaming data from Kafka and applied business logic to massage and transform and serialize raw data.
* Developed common Flink module for serializing and deserializing AVRO data by applying schema.
* Developed Spark streaming pipeline to batch real time data, detect anomalies by applying business logic and write the anomalies to HBase table.
* Leveraged cloud and GPU computing technologies for automated machine learning and analytics pipelines, such as AWS, GCP.
* Designed and developed the ETL solution using SSIS packages to implement slowly changing dimensions of Type 2 to populate current and historical data into Dimensions.
* ETL solution validated incoming data and sent notifications when jobs are done.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Implemented layered architecture for Hadoop to modularize design. Developed framework scripts to enable quick development. Designed reusable shell scripts for Hive, Sqoop, Flink and PIG jobs. Standardize error handling, logging and metadata management processes.
* Indexed processed data and created dashboards and alerts in Splunk to be utilized/ action by support teams.
* Got involved in migrating on prem Hadoop system to using GCP (Google Cloud Platform).
* Responsible for operations and support of big data Analytics platform, Splunk and Tableau visualization.
* Configured and implemented the Azure Data Factory Triggers and scheduled the Pipelines; monitored the scheduled Azure Data Factory pipelines and configured the alerts to get notification of failure pipelines.
* Used various sources to pull data into Power BI such as SQL Server, Excel, Oracle, SQL Azure etc.
* Working on Azure Data bricks to run Spark-Python Notebooks through ADF pipelines.
* Managed, developed, and designed a dashboard control panel for customers and Administrators using Tableau, Postgres SQL and RESTAPI calls.
* Designed and Developed applications using Apache Spark, Scala, Python, NIFI, S3, AWS EMR on AWS cloud to format, cleanse, validate, create schema and build data stores on S3.
* Developed CI-CD pipeline to automate build and deploy to Dev, QA, and production environments.
* Supported production jobs and developed several automated processes to handle errors and notifications. Also, tuned performance of slow jobs by improving design and configuration changes of PY Spark jobs.
* Created standard report Subscriptions and Data Driven Report Subscriptions.

**Environment: Hadoop, Map Reduce, Spark, Spark ML Lib, Tableau, SQL, Excel, PIG, Hive, AWS, Postgres SQL, Python, PY Spark, (Google cloud Platform) GCP Flink, Power BI, ADF, ETL, SSIS, Kafka, SQL Server 2012, T-SQL, CI-CD, Git, XML, Tableau.**

Client: Starbucks, Seattle Dec 2021 – July 2022

Role: Data Engineer

**Responsibilities:**

* Developed Spark SQL Scripts for data Ingestion from Oracle to Spark Clusters and relevant data joins using Spark SQL.
* Experience building distributed high-performance systems using Spark and Scala
* Experience developing Scala applications for loading/streaming data into NoSQL databases (MongoDB) and HDFS.
* Designed Distributed algorithms for identifying trends in data and processing them effectively.
* Used Spark and Scala for developing machine learning algorithms which analyses click stream data.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
* Experience in developing machine learning code using spark MLLIB
* Used Spark SQL for data pre-processing, cleaning and joining very large data sets.
* Experience in creating data lake using spark which is used for downstream applications
* Designed and Developed Scala workflows for data pull from cloud-based systems and applying transformations on it.
* Installed and configured multi-nodes on fully distributed Hadoop cluster.
* Involved in Hadoop Cluster environment administration that includes De-commissioning and commissioning nodes, cluster capacity planning, balancing, performance tuning, cluster Monitoring and Troubleshooting.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Implemented the Hadoop Name-node HA services to make the Hadoop services highly available.
* Developed the Cronjob for storing the Name-node metadata onto the NFS mount directory.
* Worked on installing Hadoop Ecosystem components such as Sqoop, Pig, Hive, Oozie, and H catalog.
* Designed Power BI data visualization utilizing cross tabs, maps, scatter plots, pie, bar and density charts.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Designed and developed ETL packages using SQL Server Integration Services (SSIS) to load the data from different source systems (SQL Server, Oracle, flat files, csv files).
* Involved in HDFS maintenance and administering it through Hadoop-Java API.
* Loaded data into the cluster from dynamically generated files using Flume and from relational database management systems using Sqoop.
* Proficient in writing Flume and Hive scripts to extract, transform and load the data into Database.
* Responsible for maintaining, managing and upgradation of Hadoop cluster connectivity and security.
* Worked on Machine Learning Algorithms Development for analyzing click stream data using Spark and Scala.
* Database migrations from Traditional Data Warehouses to Spark Clusters.
* Data Workflows and Pipelines are created for transition and analyzing trends using Spark Mllib.
* Entire Project is set up on Amazon Web Services Cloud and all the Algorithms are tuned for their best performances for better performance.
* Analyzing Streaming data and identifying important trends in data for further analysis using Spark Streaming and Storm.
* Collected and aggregated large amounts of web log data from different sources such as web servers, mobile and network devices using Apache Kafka and stored the data into HDFS for analysis.
* Experience configuring spouts and bolts in various Apache Storm topologies and validating data in the bolts.
* Effectively used data blending feature in tableau.
* Building, publishing customized interactive reports and dashboards, report scheduling using Tableau server.
* Experienced in GCP Data pro, GCS, Cloud functions, Big Query.
* Experienced in Implementing Continuous Delivery pipeline with Maven, Ant, Jenkins, and GCP.
* Used Spark Streaming to collect data from Kafka in near-real-time and perform necessary transformations and aggregation on the fly to build the common learner data model and persists the data in NoSQL store
* Populated HDFS and Cassandra with huge amounts of data using Apache Kafka.
* Batch loading of data to NOSQL storage like MongoDB.
* Implemented Spark RDD transformations, actions to migrate Map reduce algorithms.
* Used Git to check-in and checkout code changes.
* Used Jira for bug tracking

**Environment: Scala, Apache Spark, Power Bi, Azure data, ADF, AWS, GCP, SSIS, ETL, Big Query, Spark ML lib, Spark SQL, Postgres SQL, Tableau, Hive, Mongo DB, Apache Storm, Kafka, Git, Jira.**

Client: DICEUS, Wilmington Oct 2019 – Nov 2021

Role: Big Data Engineer/Spark Developer

**Responsibilities:**

* Involved in implementation of new statistical algorithms and operators on Hadoop and SQL platforms and utilized optimizations techniques, linear regressions, K-means clustering, Native Bayes and other approaches.
* Developed Spark batch job to automate creation/metadata update of external Hive table created on top of datasets residing in HDFS.
* Developed Data Serialization spark common module for converting Complex objects into sequence bits by using AVRO, PARQUET, JSON, CSV formats.
* Worked on ER Modeling, Dimensional Modeling (Star Schema, Snowflake Schema), Data warehousing and OLAP tools.
* Populated HDFS and PostgreSQL with huge amounts of data using Apache Kafka.
* Design and develop Rest API (Commerce API) which provides functionality to connect to the PostgreSQL through Java services.
* Experience in Error handling, Debugging, error logging for production support in SSIS.
* Designed Batch Audit Process in batch\shell script to monitor each ETL job along with reporting status which includes table name, start and finish time, number of rows loaded, status, etc.
* Developed Spark jobs in PY Spark to perform ETL from SQL Server to Hadoop.
* Responsible for continuous monitoring and managing Elastic MapReduce (EMR) cluster through AWS console.
* Developed scripts using PY Spark to push the data from GCP to the third-party vendors using their API framework.
* Worked on Informatica Power Center, Informatica Power Exchange for Metadata Analysis.
* Experience in GCP Data pro, GCS, Cloud functions, Cloud SQL & Big Query.
* Developed scripts using PY Spark to push the data from GCP to the third-party vendors using their API framework.
* Designed and implemented data acquisition, ingestion, Management of Hadoop infrastructure and other Analytics tools (Splunk, Tableau).
* Working knowledge of build automation and CI/CD pipelines.
* Developed python scripts to automate data ingestion pipeline for multiple data sources and deployed Apache NIFI in AWS.
* Design and develop Tableau visualizations which include preparing Dashboards using calculations, parameters, calculated fields, groups, sets and hierarchies.

**Environment: Hadoop, Map Reduce, Spark, Spark ML Lib, Tableau, SQL, Excel, PIG, Hive, AWS, Postgres SQL, Python, PY Spark, GCP, SSIS, ETL, Flink, Big Query, Kafka, SQL Server 2012, T-SQL, CI-CD, Git, XML, Tableau.**

Client: FincluS, Bangalore May 2017 – Sep 2019

Role: Data Scientist/Python Developer

**Responsibilities:**

* Performed Data Integration, Extraction, Transformation, and Load (ETL) Processes
* In preprocessing phase, used Pandas to remove or replace all the missing data and balanced the dataset with Over-sampling the minority label class and Under-sampling the majority label class to perform a well Data Cleaning Process.
* Developed models relying on Linear Regression, Multiple Regression, Decision Trees, Random Forest, Logistic Regression, Naive Bayes.
* Validated and selected models using k-fold cross validation, confusion matrices and worked on optimizing models for high recall rate and Implemented Ensemble models like Boosting and Bagging.
* Developed multi cloud strategies in better using GCP (for its PAAS).
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Provided data mining solutions on Association Rules, Classification and Clustering problems
* Used Python to identify trends and relationships between different pieces of data and drew appropriate conclusions.
* Developed Rest API to serve the data generated by prediction model to serve other customers/teams.
* Converted SAS code to python/spark-based jobs in cloud data pro/big query in GCP.
* Generated complex reports in various formats such as list reports, summary reports etc. using advanced data manipulation techniques from SAS Enterprise Guide.
* Analyzed data and generated analytical reports using SAS, MS SQL.
* Created Visual Charts, Graphs, Maps, Area Maps, Dashboards and Storytelling using Tableau.
* Develop, Maintain and support Continuous Integration framework (CI-CD) based on Jenkins.
* Implemented, tuned and tested the model on AWS EC2 with the best performing algorithm and parameters.

**Environment: AWS EC2, S3, GCP, Python (Sci Kit-Learn/NumPy/Pandas/Matplotlib), Machine Learning (Logistic Regression/Support Vector Machine/Gradient Boosting/Random Forest), Tableau, SAS, MS SQL.**

Client: Focus soft net, Hyderabad June 2015 – April 2017

Role: SQL Data Analyst

**Responsibilities:**

* Designed and created Data Marts in data warehouse database.
* Implementations of MS SQL Server Management studio 2008 to create Complex Stored Procedures and Views using T-SQL.
* Collecting the data from many resources and converting into flat text files with comma delimiter separator and importing the data to the SQL server for data manipulations.
* Responsible for deploying reports to Report Manager and Troubleshooting for any errors during the execution.
* Scheduled the reports to run on daily and weekly basis in Report Manager and emailing them to director and analyst to be reviewed in Excel Sheets.
* Created several reports for claims handling which had to be exported out to PDF formats.
* Analyzed business requirements and provided excellent and efficient solutions.

**Environment: SQL Server 2008, Microsoft Visual Studio 2008, MS Office, SSRS.**