***Manish S***

**Sr. Data Engineer**

Mobile: +1- 4698857633

Email:manish.944958@gmail.com

<https://www.linkedin.com/in/manish-s-a19395265/>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Objective:

* Experienced, ***Data Engineer*** seeking a challenging position that effectively utilize my experience in enabling the use of exceptional problem-solving, analytical, and technical skills.

**Experience Summary:**

* Having 10+ years of diversified experience in Software Design & Development. Experience as Big Data Engineer/ Python Developer solving business use cases for several supply chain clients. Experience in the field of software with expertise in backend applications.
* Extensive hands - on experience in IT industry including deployment of Hadoop Ecosystems and Google cloud computing like MapReduce, Yarn, Sqoop, Flume, Pig, Hive, Big Query, Big Table and experience on Spark, Storm, Scala, Python.
* Experience in Cloud computing on Google Cloud Platform with various technology like Dataflow, Pub/Sub, Big Query and all related tools.
* Experienced in using Agile methodologies including extreme programming, SCRUM and Test-Driven Development (TDD).
* Experienced in loading dataset into Hive for ETL (Extract, Transfer and Load) operation.
* Experience in importing and exporting data using Sqoop from Relational Database Systems to HDFS and vice - versa.
* Familiar in building DevOps pipelines for CI/CD.
* Experienced in frameworks like Flask, Django and Python packages like PySide, PyQtGraph, NumPy, MatPlot Lib.
* Experienced in developing Web Services with Python programming language.
* Excellent knowledge on Hadoop Architecture such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce programming paradigm.
* Worked with HBase to conduct quick look ups (updates, inserts and deletes) in Hadoop.
* Experience in Oozie and workflow scheduler to manage Hadoop jobs by Direct Acyclic Graph (DAG) of actions with control flows.
* Collected logs data from various sources and integrated in to HDFS using Flume.
* Extensive Experience on importing and exporting data using stream processing platforms like Flume.
* Developed Apache Spark jobs using Scala and Python for faster data processing and used Spark Core and Spark SQL libraries for querying.
* Have designed and developed ETL mapping for data collection from various data feeds using REST API.
* Extensive experience using MAVEN as a Build Tool for the building of deployable artifacts from source code.
* Proficient in usage of tools like Erwin (Data Modeler, Model Mart, navigator), ER Studio, IBM Meta Data Workbench, Oracle data modelling, profiling tool, Informatica, Oracle Forms, Reports, SQL\*Plus, Toad, Crystal Reports.
* Installed and configured OpenShift platform in managing Docker containers and Kubernetes Clusters.
* Experience in dealing with Apache Hadoop components like HDFS, MapReduce, HIVE, HBase, PIG, SQOOP, Spark and Flume Big Data and Big Data Analytics
* Expertise in relational database systems (RDBMS) such as My SQL, Oracle, MS SQL, and No SQL database systems like Hbase, MongoDB and Cassandra.
* Expertise in implementing DevOps culture through CI/CD tools like Repos, Code Deploy, Code Pipeline, GitHub.
* Experience in Microsoft Azure/Cloud Services like SQL Data Warehouse, Azure SQL Server, Azure Databricks, Azure Data Lake, Azure Blob Storage, Azure Data Factory
* Experience in developing MapReduce Programs using Apache Hadoop for analysing the big data as per the requirement.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
* Experienced in developing Web Services with Python programming language and good working experience in processing large datasets with Spark using Scala and Pyspark.
* Experience working with Amazon's AWS services like EC2, EMR, S3, KMS, Kinesis, Lambda, API gateways, IAM etc.
* Experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating data visualizations using R, SAS and Python and creating dashboards using tools like Tableau.
* Experienced in deploy to Integrate with multiple build systems and to provide an application model handling multiple projects.
* Hands on experience in data processing automation using python.
* Designed and implemented a product search service using Apache Solr.
* Flexible working Operating Systems like Unix/Linux (Centos, Redhat, Ubuntu) and Windows Environments.
* Experience with Software development tools such as JIRA, GIT, SVN.
* Experience in develop data set processes for data modelling, and Data mining. Recommend ways to improve data reliability, efficiency and quality.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Technologies** | Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume,  Impala, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper |
| **Hadoop Distribution** | Cloudera CDH, Apache, AWS, Horton Works HDP |
| **Programming Languages** | SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, Scala, Shell  Scripting, Regular Expressions |
| **Spark components** | RDD, Spark SQL (Data Frames and Dataset), and Spark Streaming |
| **Cloud Infrastructure** | AWS, MS Azure |
| **Databases** | Oracle, Teradata, My SQL, SQL Server, NoSQL Database (HBase,  MongoDB) |
| **Scripting & Query**  **Languages** | Shell scripting, Sql |
| **Version Control** | CVS, SVN and Clear Case, GIT |
| **Build Tools** | Maven, SBT |
| **Containerization Tools** | Kubernetes, Docker, Docker Swarm |
| **Reporting Tools** | Junit, Eclipse, IICS, Visual Studio, Net Beans, Azure Databricks, UNIX Eclipse,  Visual Studio, Net Beans, Junit, CI/CD, Linux, Google Shell, Unix,  Power BI, SAS and Tableau |

**Work Experience:**

**Senior Data Engineer (BigData) 02/23 - Present**

**Client - Western Union**

* Worked with the Spark for improving performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, PySpark, Impala, Tealeaf, Pair RDD's, Nifi, DevOps, Spark YARN.
* Strong Knowledge on architecture and components of Tealeaf, and efficient in working with Spark Core, SparkSQL. Designed and developed RDD Seeds using Scala and Cascading. Streaming data to Spark streaming using Kafka.
* Migrating an entire oracle database to BigQuery and using of power bi for reporting.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Loading salesforce Data every 15 min on incremental basis to BIGQUERY raw and UDM layer using SOQL, Google DataProc, GCS bucket, HIVE, Spark, Scala, Python, Gsutil And Shell Script.
* Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.
* Experience in moving data between GCP and Azure using Azure Data Factory.
* Experience in building power bi reports on Azure Analysis services for better performance.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery.
* Hands on porting the existing on-premise Hive code migration to GCP (Google Cloud Platform) BigQuery
* Design star schema in Big Query.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data.
* Developed custom Kafka producer and consumer for different publishing and subscribing to Kafka topics.
* Installed and configured Hadoop Map Reduce, HDFS, developed multiple Map Reduce jobs in java and Scala for data cleaning and preprocessing
* Implemented many Kafka ingestion jobs to consume the real time data processing and batch processing.
* Responsible for developing data pipeline with Amazon AWS to extract the data from weblogs and store in HDFS and worked extensively with Sqoop for importing metadata from Oracle.
* Good Exposure on Map Reduce programming using Java, PIG Latin Scripting and Distributed Application and HDFS.
* Involved in change implementation, monitoring and troubleshooting of AWS, Snowflake databases and cluster related issues.
* Very keen in knowing newer techno stack that Google Cloud platform (GCP) adds.
* Can work parallelly in both GCP and Azure Clouds coherently.
* Hands of experience in GCP, Big Query, GCS bucket, G - cloud function, cloud migration, cloud dataflow, Pub/suB cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver.
* Experience working in different Google Cloud Platform Technologies like Big Query, Dataflow, Dataproc, Pubsub, Airflow.
* Good experience in Tableau for Data Visualization and analysis on large data sets, drawing various conclusions and leveraged and integrated Google Cloud Storage and Big Query applications, which connected to Tableau for end user web - based dashboards and reports.
* Responsible for estimating the cluster size, monitoring and troubleshooting of the Spark databricks cluster
* Experience in using Kafka and Kafka brokers to initiate spark context and processing livestreaming.
* Migrated Map reduce jobs to Spark jobs to achieve better performance.
* Working on designing the Map Reduce and Yarn flow and writing Map Reduce scripts, performance tuning and debugging.
* Very good implementation experience of Object-Oriented concepts, Multithreading and Java/Scala
* Extracted and updated the data into HDFS using Sqoop import and export.
* Developed HIVE UDFs to incorporate external business logic into Hive script and developed join data set scripts
* Worked with various HDFS file formats like Parque, IAM, Json for serializing and deserializing.
* Implemented Cluster for NoSQL tool HBase as a part of POC to address HBase limitations.
* Used IAM to detect and stop risky identity behaviors using rules, machine learning, and other statistical algorithms.
* Responsible for developing data pipeline using Spark, Scala, Apache Kafka to ingestion the data from CSL source and store in HDFS protected folder.
* Using Spark Dataframe API in Scala for analyzing data.
* Developed end to end data processing pipelines that begin with receiving data using distributed messaging systems Kafka for persisting data into Cassandra.
* Exposure to Spark, Spark Streaming, Spark MLlib, Snowflake, Scala, and Creating the Data Frames handled in Spark with Scala.
* Written the Map Reduce programs, Hive UDFs in Java
* Developed a NIFI Workflow to pick up the data from SFTP server and send that to Kafka broker.
* Developed Oozie workflow engine to run multiple Hive, Pig, Tealeaf, Mongo DB, Git, Sqoop and Spark jobs.
* Experienced with the Scala, Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark -SQL, Pair RDD's, Spark YARN
* Good experience in using Relational databases Oracle, MY SQL, SQL Server and PostgreSQL
* Experienced Good understanding of NoSQL databases and hands on work experience in writing applications No SQL Databases HBase, Cassandra and MongoDB.

**Environment:** Hadoop (HDFS, Map Reduce), Kafka, Scala, Snowflake, AWS Services (Lambda, EMR, Auto scaling), Yarn, IAM, PostgreSql, Spark, Impala, Mongo DB, Java, Pig, DevOps, HBase, Oozie, Hue, Sqoop, Flume, Oracle, NIFI, Git.

**Business Consultant Data Engineer (BigData) 03/22– 01/23**

**Client – Toyota**

**Responsibilities:**

* Created and maintained various Shell and Python scripts for automating various processes and optimized Map Reduce code, pig scripts and performance tuning and analysis for supply chain industry.
* Set up Jenkins pipelines for CI/CD.
* Deep familiarity with Azure Security Services Azure Active Directory, RBAC, Key Vault, ADFS.
* Wrote Scripts to generate Map Reduce jobs and performed ETL procedures on the data in HDFS.
* Involved in creating HiveQL on HBase tables and importing efficient work order data into Hive tables
* Used Pig Latin at client side cluster and HiveQL at server side cluster.
* Experience in importing and exporting Terabytes of data between HDFS and Relational Database Systems using Sqoop.
* Created Pipelines in ADF (Azure Data Factory)using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Installed and configured Hive, Pig, Sqoop, Flume and Oozie on the Hadoop cluster.
* Extensive usage of Azure Portal, Azure PowerShell, Storage Accounts, Certificates and Azure Data Management.
* Develop and maintain operating procedures and support documentation for ETL packages and the SSIS infrastructure for vehicle supply chain teams.
* Implemented data Ingestion and handling clusters in real time processing using Apache Kafka.
* Supported Map Reduce Programs those are running on the cluster.
* Developed Spark scripts by using Scala shell commands as per the requirement.
* Involved in converting Map Reduce programs into Spark transformations using Spark RDD's using Scala and Python.
* Imported documents into HDFS, HBase and creating HAR files.
* Developed Oozie workflow schedulers to run multiple Hive and Pig jobs that run independently with time and data availability.
* Ingested data into HDFS using Sqoop and scheduled an incremental load to HDFS.
* Implemented OLAP multi-dimensional cube functionality using Azure SQL Data Warehouse.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.
* Developed Json Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Cosmos Activity.
* Used Zookeeper to provide coordination services to the cluster.
* Wrote python scripts to parse XML and JSON data and load the data into database.
* Installed Oozie workflow engine to run multiple Hive.
* Experience in using the Docker container system with the Kubernetes integration
* Wrote AZURE POWERSHELL scripts to copy or move data from local file system to HDFS Blob storage.
* Involved in installation, configuration, supporting and managing Hadoop clusters, Hadoop cluster administration.
* Utilized Spark, Scala, Hadoop, HBase, Kafka, Spark Streaming, MLLib, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Worked on ETL using Spark, Spark Streaming, Hive, HBase, and Oozie on Hadoop.
* Implemented Apache Drill on Hadoop to join data from SQL and No SQL databases and store it i Configured Spark streaming to receive real time data from Kafka and store the stream data to HDFS using Scala.
* Developed Spark Applications by using Scala, Python and Implemented Apache Spark data processing Project to handle data from various RDBMS and Streaming sources.
* Developed real time SLA monitoring dashboards in Tableau for the Kafka messages load in Sap HANA
* Performed optimizing Map Reduce Programs using combiners and custom counters for delivering the best results.
* Automated jobs using different triggers (Event, Scheduled and Tumbling) in ADF (Azure Data Factory) and Created pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Data bricks.
* Working on Azure Data bricks to run Spark-Python Notebooks through ADF pipelines.
* Experience with scalable architectures using Azure App Service, API management, serverless technologies
* Have good knowledge on NoSQL databases like HBase, Cassandra and MongoDB.
* Synchronizing both the unstructured and structured data using Pig and Hive on business prospectus.
* Used Zookeeper to provide coordination services to the cluster. Experienced in managing and reviewing Hadoop log files.
* Using Data bricks utilities called widgets to pass parameters on run time from ADF to Data bricks.
* Implemented data access jobs through Pig, Hive, Tez, Solr, Accumulo, Hbase, and Storm.
* Good Experience on importing and exporting the data from HDFS and Hive into Relational Database Systems like MySQL and vice versa using Sqoop for data modelling.
* Experience in migrating the data using Sqoop from HDFS and Hive to Relational Database System and vice-versa according to supply chain client's requirement.
* Enabled Python scripts to explode, parse and de-dupe JSON from Kafka and land in HDFS.
* Involved in creating Hive tables, loading with data and writing hive queries that will run internally in map reduce way.

**Environment:** Hadoop, HDFS, MapReduce, Kafka, Spark, Azure, Data Bricks, Python, Scala, Pig, Hive, PL/SQL, SQL, Oracle 12c, Mongo DB, Jenkins, Docker, Kubernetes, Git

**Data Engineer 06/20 – 12/21**

**Client - United Airlines**

* Wrote Spark applications and mentored other team members on the perks of spark.
* Developed spark applications in python (PySpark) on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
* Worked on setting up and configuring AWS's EMR Clusters and Used Amazon IAM to grant fine-grained access to AWS resources to supply chain users.
* Worked on AWS Lambda functions in python for AWS Lambda which invokes python scripts to perform various transformations and analytics on large data sets in EMR clusters.
* Developed Apache Spark applications by using spark for data processing from various streaming sources.
* Developed Java Map Reduce programs for the analysis of sample log file stored in cluster.
* Implemented Data Quality in ETL Tool Talend and having good knowledge in Data modelling and Warehousing
* Installed application on AWS EC2 instances and configured the storage on S3 buckets.
* Evaluating supply chain client needs and translating their business requirement to functional specifications thereby onboarding them onto Hadoop ecosystem.
* Used the AWS-CLI to suspend an AWS Lambda function. Used AWS CLI to automate backups of ephemeral data-stores to S3 buckets, EBS.
* Expert in migrating data from various systems in Salesforce CRM using ETL tools.
* Created functions and assigned roles in AWS Lambda to run python scripts, and AWS Lambda using java to perform event driven processing. Created Lambda jobs and configured Roles using AWS CLI.
* Stored data in AWS S3 like HDFS and performed EMR programs on data stored.
* Experience in data ingestions techniques for batch and stream processing using AWS Batch, AWS Kinesis, AWS Data Pipeline
* Wrote complex logic implementations using Spark to process data present in MapR DB and Hive.
* Implemented Spark using Scala, Python and utilizing Data frames and Spark SQL API for faster processing of data.
* Experience in developing customized UDF’s in Python to extend Hive and Pig Latin functionality.
* Building data pipeline ETLs for data movement to S3, tan to Redshift.
* The near real time reporting was achieved by an event-based processing approach adoption instead of micro-batching to deal with data coming from Kafka for flight supply chain.
* Extensive usage of Spark for data streaming and data transformation for real time analytics.
* Queried and analyzed data from DataStax Cassandra for quick searching, sorting and grouping.
* Experienced in writing real-time processing and core jobs using Spark Streaming with Kafka as a data pipeline system.
* Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket.
* Have written applications using Spring boot that reads data from Kafka and writes it to MapR DB (MapR version of HBase).
* Worked extensively on AWS Components such as Elastic Map Reduce (EMR)
* Supported Map Reduce Programs those are running on the cluster. Involved in loading data from UNIX file system to HDFS.
* Written AWS Lambda code in Python for nested Json files, converting, comparing, sorting etc.
* Setup and benchmarked Hadoop/Hbase clusters for internal use.
* Worked on reading and writing multiple data formats like JSON, ORC, Parquet on HDFS using PySpark
* Experienced in working with various kinds of data sources such as Teradata and Oracle. Successfully loaded files to HDFS from Teradata, and load loaded from HDFS to hive and impala.
* Optimization of Hive queries using best practices and right parameters and using technologies like Hadoop, YARN, Python, PySpark.
* Have written HiveQL scripts to populate table and brought data from various systems using Sqoop.
* Used DataStax Cassandra along with Pentaho for reporting.
* Created data pipeline for different events of ingestion, aggregation and load consumer response data in AWS S3 bucket into Hive external tables in HDFS location to serve as feed for Tableau dashboards
* Installed and configured Hive and written Hive UDFs and Used Map Reduce and Junit for unit testing.
* Installed and configured Apache Hadoop to test the maintenance of log files in Hadoop cluster.
* Built a dashboard of all the YARN applications running on the cluster using YARN API.
* Have written applications that produced data to Kafka and also consumed data from it.
* Used Scala to convert Hive/SQL queries into RDD transformations in Apache Spark.
* Implemented Spark solutions to generate reports, fetch and load data in Hive.
* Used Yarn Architecture and Map reduce in the development cluster for POC
* Data Import and Export from various sources through Script and Sqoop
* Involved in various NOSQL databases like Hbase, Cassandra in implementing and integration
* Developed Java Map Reduce programs for the analysis of sample log file stored in cluster.

**Environment:** Hadoop, HDFS, AWS, Apache Spark, Pyspark, Impala, MapReduce, Hive, Kafka, HBase, Sqoop, Python, Spark, Yarn, Spring boot, No Sql, Hbase, Cassandra, Relational Databases, Oracle 12c, SQL Server, Linux, Unix.

**Data Engineer         09/17 – 03/20**

**Client - Honeywell**

* Installed and configured Hadoop Map Reduce, HDFS, developed multiple MapReduce jobs in java and Scala for data cleaning and pre-processing.
* Used Yarn Architecture and MapReduce in the development cluster for POC.
* Used existing Deal Model in Python to inherit and create object data structure for regulatory reporting.
* Designed and implemented a product search service using Apache Solr/Lucene.
* Integrated Cassandra as a distributed persistent metadata store to provide metadata resolution for network entities on the network supply chain
* Worked in installing cluster, commissioning & decommissioning of Data nodes, Name node recovery, capacity planning, and slots configuration.
* Installed and configured Hive and written Hive UDFs and Used Map Reduce and Junit for unit testing.
* Experienced in running query-using Impala and used BI tools to run adhoc queries directly on Hadoop.
* Transform and load (ETL) existing data from the legacy systems and other sources into AWS S3.
* Queried and analyzed data from DataStax Cassandra for quick searching, sorting and grouping.
* Experienced in working with various kinds of data sources such as Teradata and Oracle. Successfully loaded files to HDFS from Teradata, and load loaded from HDFS to hive and impala.
* Airflow to schedule ETL jobs and IICS and Glue and Athena to extract the data from AWS data warehouse
* Controlled backend logic using Python.
* Implemented Kafka consumers to move data from Kafka partitions into Cassandra for near real-time analysis.
* Involved in various NOSQL databases like Hbase, Cassandra in data modelling ,implementing and integration.
* Experienced in Importing and exporting data into HDFS and Hive using Sqoop.
* Participated in development/implementation of Cloudera Hadoop environment.

**Environment:** CDH, MapReduce, Scala, Kafka, AWS, spark, Solr, HDFS, Hive, pig, Impala, Cassandra, Java, SQL, Tableau, PIG, Zookeeper, Pentaho, Sqoop, Python, Teradata, CentOS.

**Client: Veriton Software, Bangalore, India 10/15 – 12/16**

**Role: Data Engineer**

* Installed and configured Hadoop Map Reduce, HDFS, developed multiple MapReduce jobs in java and
* Scala for data cleaning and pre-processing.
* Used Yarn Architecture and MapReduce in the development cluster for POC.
* Used existing Deal Model in Python to inherit and create object data structure for regulatory reporting.
* Designed and implemented a product search service using Apache Solr/Lucene.
* Integrated Cassandra as a distributed persistent metadata store to provide metadata resolution for network entities on the network
* Worked in installing cluster, commissioning & decommissioning of Data nodes, Name node recovery, capacity planning, and slots configuration.
* Installed and configured Hive and written Hive UDFs and Used Map Reduce and Junit for unit testing.
* Experienced in running query-using Impala and used BI tools to run adhoc queries directly on Hadoop.
* Transform and load (ETL) existing data IICS from the legacy systems and other sources into AWS S3.
* Queried and analyzed data from DataStax Cassandra for quick searching, sorting and grouping.
* Experienced in working with various kinds of data sources such as Teradata and Oracle. Successfully loaded files to HDFS from Teradata, and load loaded from HDFS to hive and impala.
* Airflow to schedule ETL jobs and Glue and Athena to extract the data from AWS data warehouse
* Controlled backend logic using Python.
* Implemented Kafka consumers to move data from Kafka partitions into Cassandra for near real-time analysis.
* Involved in various NOSQL databases like Hbase, Cassandra in implementing and integration.
* Experienced in Importing and exporting data into HDFS and Hive using Sqoop.

Participated in development/implementation of Cloudera Hadoop environment.

**Environment**: IICS, CDH, MapReduce, Scala, Kafka, AWS, spark, Solr, HDFS, Hive, pig, Impala, Cassandra, Java, SQL, Tableau, PIG, Zookeeper, Pentaho, Sqoop, Python, Teradata, CentOS.

**Data Hadoop Analyst**

**GS Industries         03/12 – 09/15**

* Designed, implemented, and deployed within a customer’s existing Hadoop / Cassandra cluster a series of custom parallel algorithms for various customer defined metrics and unsupervised learning models.
* Installed and configured Hive, Pig, Sqoop, Flume and Oozie on the Hadoop cluster.
* Installed Oozie workflow engine to run multiple Hive and Pig Jobs.
* Implemented Avro and parquet data formats for apache Hive computations to handle custom business requirements.
* Developed Simple to complex Map/reduce Jobs using Hive and Pig
* Developed Map Reduce Programs for data analysis and data cleaning.
* Performed data cleansing, enrichment, mapping tasks and automated data validation processes to ensure meaningful and accurate data was reported efficiently.
* Implemented Avro and parquet data formats for apache Hive computations to handle custom business requirements.
* Developed Simple to complex Map/reduce Jobs using Hive and Pig
* Developed Map Reduce Programs for data analysis and data cleaning.
* Performed data cleansing, enrichment, mapping tasks and automated data validation processes to ensure meaningful and accurate data was reported efficiently.
* Implemented Apache PIG scripts to load data from and to store data into Hive.
* Extensively used SSIS transformations such as Lookup, Derived column, Data conversion, Aggregate, Conditional split, SQL task, Script task and Send Mail task etc.
* Implemented Apache PIG scripts to load data from and to store data into Hive.
* Extensively used SSIS transformations such as Lookup, Derived column, Data conversion, Aggregate, Conditional split, SQL task, Script task and Send Mail task etc.

**Environment:** Hive, Hadoop, Cassandra, Pig, Sqoop, Ooze, Hive, Scala. Python, MS Office.