**Nandakumar Konda**

 **Nanda.Tech90@gmail.com**

 **+1(740) 490-2525**

**Professional Summary:**

* **Over 10+ years of** experience in Big Data Analytics, Hadoop, Java, Database Administration, GCP and Software development expertise.
* Strong hands-on experience in Hadoop Framework and its ecosystem including **HDFS Architecture, MapReduce Programming, Hive, Pig, Sqoop, Hbase, Zookeeper, Couchbase, Storm, Solr, Oozie, Spark, Scala, Airflow, Flume, Strom and Kafka.**
* Excellent knowledge on Hadoop Architecture and ecosystems such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce programming** paradigm.
* Experience in analyzing data using **HIVEQL** and **Pig Latin** and custom **Map Reduce** programs in **Java** and **scala**.
* Experience in strong and analyzing data using **HiveQL, Pig Latin, HBase and custom Map Reduce programs in Java.**
* Experience in importing and exporting data into **HDFS** and **Hive** using **Sqoop.**
* Integrated different data sources, data wrangling: cleaning, **transforming**, **merging** and **reshaping** **data** sets by writing **Python** **scripts**.
* Good knowledge on Amazon AWS concepts like EMR & EC2 web services which provides fast and efficient processing of Big Data.
* Hands on experience on Unified Data Analytics with **Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks,** Delta Lake with Python**,** Delta Lake with Spark SQL.
* Experienced in Cloud computing on **Google Cloud Platform** with various technologies like Dataflow, Big Query and related tools.
* Work with client teams to design and implement modern, scalable data solutions using a range of new and emerging technologies from the Google Cloud Platform.
* Hands on experience in installing, configuring Cloudera's Apache Hadoop ecosystem components like
* **Flume-ng**, **Hbase**, **Zoo Keeper**, **Oozie**, **Airflow** **Hive**, **Spark**, **Storm, Sqoop, Kafka, Hue, Pig, Hue** with **CDH3&4** Clusters
* Architected, Designed and maintained high performing **ELT/ETL Processes**.
* Skilled in managing and reviewing **Hadoop log files**.
* Experienced in loading data to Hive partitions and creating buckets in **Hive**
* Experienced in configuring **Flume** to stream data into HDFS.
* **Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.**
* Experienced in real-time Big Data solutions using **Hbase**, handling billions of records.
* Processing this data using Spark Streaming API with Scala.
* Familiarity with distributed coordination system **Zookeeper**.
* Be required to build and deliver Data solutions using GCP products and offerings.
* Efficient in writing MapReduce programs and using **Apache** **Hadoop API** for analyzing the structures and unstructured data
* Involved in designing and deploying a multitude application utilizing the entire AWS stack (Including EC2, RDS, VPC, IAM) focusing on high-availability, fault tolerance and auto-scaling.
* Experienced in implementing unified data platforms using **Kafka** producers/ consumers, implement pre-processing using **storm** topologies.
* Good knowledge on building Apache spark applications using **Scala**.
* Big Query storage is automatically replicated across multiple locations to provide high availability Execute bigquery based on business requirements.
* Hands on experience in setting up workflow using Apache Airflow and Oozie workflow engine for managing and scheduling Hadoop Jobs.
* Utilize GCP services with focus on big data Architect /analytics / enterprise data warehouse and business intelligence solutions to ensure optimal architecture, scalability, flexibility, availability, performance, and to provide meaningful and valuable information for better decision-making.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in developing and designing **POCs** using Scala and deployed on the Yarn cluster, compared the performance of **Spark** with **Hive** and **SQL/Teradata**.
* Potential experience in (SDLC) Analysis, Design, Development, Integration and Testing in diversified areas of Client-Server/Enterprise applications using **Java, J2EE technologies**.
* Done **Administration, installing, upgrading** and **managing** distributions of **Cassandra**.
* Strong database development skills using Database servers like **Oracle, IBM DB2, My SQL** and hands on experience with **SQL, PL/SQL.** Extensive experience of backend database programming in oracle environment using PL/SQL with tools such as **TOAD.**
* Have a very good understanding and worked with relational databases like **MySQL**, **Oracle** and **NoSQL** **databases** like **Hbase**, **Mongo** **DB**, **Couchbase** and **Cassandra**.
* Good work experience on **JAVA, JDBC, Servlets, JSP.**
* Proficient in **Java, J2EE, JDBC, Collections, Servlets, JSP, Struts, Spring, Hibernate, JAXB, JSON, XML, XSLT, XSD, JMS, WSDL, WADL, REST, SOAP Web services, CXF, Groovy, Grails, Jersey, Gradle and Eclipse Link.**
* Good knowledge in performance troubleshooting and tunning **Cassandra clusters** and understanding of **Cassandra Data Modeling** based on applications.
* Good knowledge in integration of various data sources like **RDBMS, Spreadsheets, Text files, JSON and XML files**.
* Skilled in developing applications in **Python** language for multiple platforms familiarity with process and **Python** software development

**Education:**

* MS in Information Assurance from Wilmington university in DE, 2018
* Bachelors from Osmania University, 2010

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Eco System** | Hadoop 2.1, HDFS, MapReduce, PIG 0.8, Hive0.13, Hbase 0.94, Sqoop 1.4.4, Zookeeper 3.4.5, Storm, Yarn, Spark Streaming, Spark SQL, Kafka, Scala, Cloudera CDH3, CDH4, Hortonworks, Azure Databricks, Oozie, Apache Airflow, Flume, Impala, Talend, Tableau/Qlickview |
| **Hadoop management & Security** | Hortonworks Ambari, Cloudera Manager, Kafka, Data Proc |
| **NoSQL Databases** | MongoDB, Hbase, Redis, Couchbase and Cassandra |
| **Web Technologies** | DHTML, HTML, XHTML, XML, XSL (XSLT, XPATH), XSD, CSS, JavaScript, Servlets, SOAP, Amazon AWS, GCP, GCS |
| **Server-Side Scripting** | UNIX Shell Scripting |
| **Database** | Oracle 11g/10g/9i/8i, MS SQL Server 2012/2008, DB2 v8.1, MySQL, Teradata, Snowflake |
| **Programming Languages** | Java, J2EE, JSTL, JDBC 3.0/2.1, JSP 1.2/1.1 |
| **Scripting Languages** | Python, Perl, Shell Scripting, JavaScript, Scala |
| **OS/Platforms** | Windows7/2008/Vista/2003/XP/2000/NT, Macintosh, RedHat EL 4.x/5.x/Linux (All major distributions, mainly Centos and Ubuntu), Unix |
| **Client side** | JavaScript, CSS, HTML, JQuery, BigQuery |
| **Build tools** | Maven and ANT |
| **Methodologies** | Agile, UML, Design Patterns, SDLC |
| **Tools** | FileZilla, Putty, TOAD SQL Client, MySQL Workbench, ETL, DWH, JUnit, SQL Oracle Developer |
| **Office Tools** | MS Office - Excel, Word, PowerPoint |

**Professional Experience:**

**Elevance Health (Anthem) Dec’21 – Till Date**

**Sr. Data Engineer (GCP)**

Anthem, Inc., is a provider of health insurance in the United States. It is the largest for-profit managed health care company in the Blue Cross Blue Shield Association.

**Responsibilities:**

* Working in an Agile team to deliver and support required business objectives by using Java, Python and shell scripting and other related technologies to acquire, ingest, transform and publish data both to and from **Hadoop Ecosystem.**
* Loaded data into the cluster from dynamically generated files using **Flume** and from relational database management systems using **Sqoop**.
* Used **Flume** to collect, aggregate and store the web log data onto **HDFS**.
* Performed Data Cleansing using **Python** and loaded into the target tables.
* Logical implementation and interaction with **HBASE**.
* Used **Scala** to store streaming data to HDFS and to implement Spark for faster processing of data.
* Integrating user data from **Cassandra** to **HDFS**. Integrating Cassandra with Storm for real time user attributes look up.
* Developed **Kafka** consumer's **API** in **scala** for consuming data from Kafka topics
* Developed 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.
* Performed Sqoop Incremental imports by using **Oozie** based on every day.
* Installed and configured Hadoop MapReduce, HDFS, developed MapReduce jobs in Java for data cleaning and pre-processing.
* Supported middleware environments like **Apache/Tomcat** running on **Redhat Linux**.
* Involved in using **HCATALOG** to access Hive table metadata from MapReduce or Pig code.
* Created **Pig** scripts to transform the HDFS data and loaded the data into **Hive** external table.
* Worked on large-scale Hadoop **YARN** cluster for distributed data processing and analysis using **Connectors, Spark core, Spark SQL, Sqoop, Hive** and **NoSQL databases**.
* Implemented Spark Scripts using **Scala, Spark SQL** to access hive tables into spark for faster processing of data.
* Involved in converting Hive/SQL queries into Spark transformations using **Spark RDDs, Python** and **Scala.**
* Performed advanced procedure like text analytics and processing, using the in-memory computing capabilities of Spark using **Scala.**
* Implemented **Spark RDD** transformations, actions to implement business analysis.
* Connected to HDFS using Pentaho Kettle to read data from hive tables and perform analysis.
* Worked on **Spark Streaming** and **Spark SQL** to run sophisticated applications on Hadoop.
* Used **Oozie** and **Oozie coordinators** to deploy end to end processing pipelines and scheduling the workflows.
* Create and Share reports and dashboards based on GCS data.
* Stores and retrieve wide range data which stored from Google cloud storage (GCS)
* Worked on concept of quorum with **Kafka** and **Zookeeper**.
* Created and maintained technical documentation for launching Hadoop Clusters and executing **Script.**
* Trouble shoot and take corrective action for problems involving **Control-M** dependencies using conditions/resource manual conditions utility.
* Part of developing a fully operational production grade large scale data solution on Snowflake Data Warehouse.
* Maintain all existing **Control-M scheduling** jobs definitions.
* Daily basis Extracting, loading, or transforming data that needs to be Automated using airflow in Apache.
* Experience on cloud based technologies like GCP and Amazon AWS concepts like EMR & EC2 web services which provides fast and efficient processing of Big Data.
* Configured and deployed instances on GCP environments.
* Interact with **Cloverleaf** team and provide them Extracts based on their requirements.
* Fixing production issues and providing error free solution.
* Developed a POC for project migration from on prem Hadoop MapR system to GCP/Snowflake
* Co-ordination with onsite/offshore team members on daily basis.
* Interact with Data Architects follow the enterprise standers on new requests for Data Models on SAP PowerDesigner.

**Environment:** Hadoop, CDH 4, CDH 5, Scala, MapReduce, HDFS, Hive, Sqoop, HBASE, Flume, Spark SQL, Snowflake, Databricks, Redhat Linux, Teradata, Microsoft SQL server Management Studio, GCP, GCS, Spark-Streaming, CntrlM Scheduler, MapR, Python, Microsoft Visual Studio. SAP PowerDesigner for Data Modelling, UNIX Shell Scripting and Cassandra.

**blue cross blue shield, durham, nc. Nov 2019 – Nov 2021**

**Position: Senior Hadoop Developer**

**Description:**

Blue Cross Blue Shield Association is a federation of 36 separate United States health insurance companies that provide health insurance in the United States to more than 106 million people. The BCBS manages communications between its members and the operating policies required to be a licensee of the trademarks.

**Job Responsibilities:**

* Extract data from multiple sources, integrate disparate data into a common data model, and integrate data into a target database, application, or file using efficient programming processes
* Document, and test moderate data systems that bring together data from disparate sources, making it available to data scientists, and other users using scripting and/or programming languages.
* Experience working on Google Cloud Big data Technologies like Data Proc, Data Flow, Big Query and GCP Storage, and having knowledge on pub sub.
* Write and refine code to ensure performance and reliability of data extraction and processing
* Participate in requirements gathering sessions with business and technical staff to distill technical requirement from business requests.
* Be required to build and deliver Data solutions using GCP products and offerings.
* Involved in the process designing Google cloud architecture
* Implemented Big Data Analytics and advanced data science techniques to identify trends, patterns and discrepancies on petabytes of data by using Azure Databricks, Hive, Hadoop, Python, PySpark, Spark SQL, MapReduce and Azure machine learning
* Develop SQL queries to extract data for analysis and model construction
* Performed Linux system administration on production and development servers (Redhat Linux, CentOS and other Unix utilities.)
* Used SparkAPI over Hortonworks Hadoop YARN to perform analytics on data in Hive.
* Own delivery of moderately sized data engineering projects
* Define and implement integrated data models, allowing integration of data from multiple sources
* Design and develop scalable, efficient data pipeline processes to handle data ingestion, cleansing, transformation, integration, and validation required to provide access to prepared data sets to analysts and data scientists
* Ensure performance and reliability of data processes
* Define and implement data stores based on system requirements and consumer requirements
* Document and test data processes including performance of through data validation and verification
* Collaborate with cross functional team to resolve data quality and operational issues and ensure timely delivery of products
* Develop and implement scripts for database and data process maintenance, monitoring, and performance tuning
* Analyze and evaluate databases in order to identify and recommend improvements and optimization
* Design eye-catching visualizations to convey information to users

**NTT DATA/Texas Department of Transportation, Austin TX Nov 2018 – Oct 2019**

**Senior Hadoop Data Engineer**

**Description:** NTT DATA is basically an IT services company. It is a SAP gold partner in implementing SAP systems. There is one more company in India called **NTT DATA** GDS (Global Delivery Services) which is services company.  Its business areas are in national and local governments, financial, and telecommunication sectors.

**Responsibilities:**

* Hadoop development and implementation (Environment - HDFS, Hbase, Spark, Kafka, Ozie, Apache Airflow, Scoop, Flume, Kerberos, Oracle ASO, MySQL)
* Loading from disparate data sets using Hadoop stack of ingestion and workflow tools
* Pre-processing using Hive and Pig.
* Worked on Data Lake, data ingestion and processing pipelines.
* Designing, building, installing, configuring and supporting Hadoop.
* Translate complex functional and technical requirements into detailed design.
* Perform analysis of vast data stores and uncover insights.
* Maintain security and data privacy.
* Managing and deploying HBase.
* Being a part of a POC effort to help build new Hadoop clusters.
* Test prototypes and oversee handover to operational teams.
* Propose best practices/standards.
* Configure and implementation of Data Marts in Hadoop platform
* Involved in loading data from Teradata, Oracle database into HDFS using **Sqoop queries**.
* Worked on setting up Kafka for streaming data and monitoring for the Kafka Cluster.
* Responsible for importing log files from various sources into **HDFS** using **Flume**.
* Imported data using **Sqoop** to load data from **MySQL to HDFS** on regular basis.
* Worked on **shell scripting** in **Linux** and the Cluster. Used shell scripts to run hive queries from beeline.
* Developed Scripts and automated data management from end to end and sync up between all the clusters.
* Used AWS EMR Spark cluster and Cloud Dataflow on GCP to compare the efficiency of a POC on a developed pipeline
* Worked with **Hue** GUI in scheduling jobs with ease and File browsing, Job browsing, Metastore management.
* Involved in creating **Hive Tables**, loading with data and writing **Hive queries** which will invoke and run MapReduce jobs in the backend.
* Developed workflow in **Oozie or Airflow** to automate the tasks of loading the data into HDFS and processing with **Sqoop** and **Hive**.
* Developed **Spark** jobs using **Scala** in test environment for faster data processing and used **Spark SQL** for querying.
* Monitoring the daily schedule jobs in Apache Airflow and find out solutions for failure jobs and provide immediate solutions.
* Created Partitions, Buckets based on State to further process using Bucket based **Hive joins**.
* Loaded multiple **NOSQL** **databases** including **MongoDB**, **PostgreSQL**, **Couchbase**, **HBase** and **Cassandra**.
* Actively participated in software development lifecycle (scope, design, implement, deploy, test), including design and code reviews, test development, test automation.
* Setting up Snowflake connections through private link from **AWS** **EC2** and AWS EMR to secure data transfers between application and database.
* Used **Zookeeper** for providing coordinating services to the cluster.

**Environment:** Hadoop, HDFS, Pig, Hive, Map Reduce, Spark, GCP, GCS,Sqoop, Kafka, Oozie, and Big Data, Python, Apache Java (jdk1.6), Data tax, Flat files, MySQL, Toad, Windows NT, LINUX, Cassandra UNIX, SVN, Data proc,Hortonworks Cassandra, AVRO Files, SQL, ETL, DWH, Cloudera Manager, Talend, Scala, MongoDB.

**First Data, Atlanta, GA May 2018 – Oct 2018**

**Position: Hadoop Developer**

**Description:** First Data Corporation provides electronic commerce solutions for merchants, financial institutions, and card issuers worldwide. It operates through three segments: Global Business Solutions (GBS), Global Financial Solutions (GFS), and Network & Security Solutions (NSS).

**Responsibilities:**

* Responsible for building **scalable distributed data** solutions using Hadoop.
* Developed multiple **Map Reduce** jobs in java for data cleaning and preprocessing.
* Developed Map Reduce pipeline jobs to process the data and create necessary HFiles.
* Involved in loading the created HFiles into Hbase for faster access of large customer base without taking Performance hit.
* Worked in AWS environment for development and deployment of Custom Hadoop Applications.
* Involved in creation and designing of data ingest pipelines using technologies such as Apache Strom and Kafka.
* Developed Spark scripts by using Scala shell commands as per the requirement.
* Implemented discretization and binning, data wrangling: cleaning, transforming, merging and reshaping data frames using Python.
* Created **Hbase** tables to store various data formats of PII data coming from different portfolios.
* Collecting and aggregating large amounts of log data using **Apache Flume** and staging data in HDFS for further analysis
* Involved in managing and reviewing Hadoop log files.
* Responsible to manage data coming from different sources.
* Involved in creating Pig tables, loading with data and writing **Pig Latin** queries which will run internally in Map Reduce way.
* Experienced in Using **Pig** as ETL tool to do Transformations, even joins and some pre-aggregations before storing the data onto HDFS.
* Transferred the data using Informatica tool from AWS S3 to AWS Redshift. Involved in file movements between HDFS and AWS S3.
* Create a complete processing engine, based on **Hortonworks’** distribution, enhanced to performance.
* Provide batch processing solution to certain unstructured and large volume of data by using **Hadoop Map Reduce framework.**
* Developed Spark code to using Scala and Spark-SQL for faster processing and testing.
* Used **Avro**Serdes to handle **Avro** Format Data in **Hive** and **Impala**.
* Involved in converting **Hive/SQL** queries into **Spark** transformations using Spark RDDs and **Scala**.
* Worked on python files to load the data from csv, json, MySQL, hive files to Neo4j Graphical database.
* Handled **Administration**, installing, upgrading and managing distributions of **Cassandra**.
* Assisted in performing unit testing of **Map Reduce** jobs using MRUnit.
* Assisted in exporting data into **Cassandra** and writing column families to provide fast listing outputs.
* Used **Oozie Scheduler** system to automate the pipeline workflow and orchestrate the map reduce jobs that extract the data on a timely manner.
* Experience in integrating **Apache Kafka**with **Apache Storm** and created Storm data pipelines for real time processing.
* Experienced in working with Spark eco system using **SCALA** and **HIVE** Queries on different data formats like Text file and parquet.
* Exposure on usage of **Apache Kafka** develop data pipeline of logs as a stream of messages using producers and consumers.
* Worked with **Hue** GUI in scheduling jobs with ease and File browsing, Job browsing, Metastore management.
* Worked with **Talend** on a POC for integration of data from the data lake.
* Highly involved in development/implementation of Cassandra environment.
* Involved in story-driven **agile development methodology** and actively participated in daily scrum meetings.

**Environment:** Hadoop, Map Reduce, Big Query, HDFS, Python, Hive, Spark, Hue, Pig, Sqoop, Kafka, GCP, GCS, AWS,Avro, HBase, Oozie, Cassandra, Impala, Zookeeper, Talend, Teradata, Oracle 11g/10g, Python, Java (jdk1.6), Scala, UNIX, SVN, Hortonworks, Maven.

**Abbott Laboratories, IL Jan 2017 – May 2018**

**Position: Hadoop Developer**

**Description:** Abbott Laboratories is an American health care company with headquarters in Lake Bluff, Illinois, United States. The company was founded by Chicago physician Wallace Calvin Abbott in 1888 to formulate known drugs; it eventually grew to also sell research-based drugs, medical devices, diagnostics, and nutritional products.

**Responsibilities:**

* Responsible to manage data coming from different sources and involved in **HDFS** maintenance and loading of structured and unstructured data.
* Developed data pipeline using **Flume, Sqoop, Pig**and MapReduce to ingest behavioral data into HDFS for analysis.
* Be required to build and deliver Data solutions using GCP products and offerings.
* Extracted files from **MongoDB** through **Sqoop** and placed in **HDFS** and processed.
* Created customized BI tool for manager team that perform Query analytics using **HiveQL.**
* Created Partitions, Buckets based on State to further process using Bucket based **Hive** joins.
* Experienced in using **Kafka** as a data pipeline between **JMS** and **Spark** Streaming Applications.
* Created storage with Amazon S3 for storing data. Worked on transferring data from Kafka topic into AWS S3 storage.
* Worked on python files to load the data from **csv, json, mysql, hive** files to **Neo4j** Graphical database.
* Estimated the hardware requirements for Name Node and Data Nodes & planning the cluster.
* Created Hive Generic **UDF'**s, **UDAF**'s, **UDTF**'s in java to process business logic that varies based on policy.
* Moved Relational Database data using **Sqoop** into **Hive** Dynamic partition tables using staging tables.
* Consolidating customer data from Lending, Insurance, Trading and Billing systems into data warehouse and mart subsequently for business intelligence reporting.
* Optimizing the **Hive** queries using Partitioning and Bucketing techniques, for controlling the data distribution.
* Experienced on Loading streaming data into HDFS using Kafka messaging system.
* Used the Spark -**Cassandra** Connector to load data to and from **Cassandra**.
* Worked with **NoSQL** database **Hbase** to create tables and store data.
* Proficient in querying Hbase using Impala.
* Worked on custom **Pig** Loaders and storage classes to work with variety of data formats such as JSON and **XM**L file formats.
* Used **Pig** as ETL tool to do Transformations, even joins and some pre-aggregations before storing the data onto **HDFS**.
* Design technical solution for real-time analytics using **Kafka** and **Hbase.**
* Created **UDF's** to store specialized data structures in **HBase** and **Cassandra**.
* Used **Pig** as ETL tool to do transformations, event joins, filter and some pre-aggregations.
* Collaborated with Business users for requirement gathering for building **Tableau** reports per business needs.
* Experience in Upgrading **Apache Ambari, CDH** and **HDP** Cluster.
* Configured and Maintained different topologies in storm cluster and deployed them on regular basis.
* Imported structured data, tables into **Hbase.**
* Involved in Backup, HA, and DR planning of applications in AWS.
* Used Impala to read, write and query the **Hadoop** **data** in HDFS from **HBase** or **Cassandra**.
* Experienced with different kind of compression techniques like LZO, GZip, and Snappy.
* Used **Oozie** workflow engine to manage interdependent Hadoop jobs and to automate several types of Hadoop jobs such as Java **map-reduce Hive, Pig, and Sqoop**.
* Used AWS Patch Manager to select and deploy operating system and software patches across EC2 instances.
* Created Data Pipeline of Map Reduce programs using Chained Mappers.
* Configuring **Spark Streaming** to receive real time data from the **Kafka** and Store the stream data to **HDFS.**
* Implemented **Spark RDD** transformations, actions to migrate Map reduce algorithms.
* Set-up configured and optimized the Cassandra cluster. Developed real-time java-based application to work along with the Cassandra database.
* Implemented Optimized join base by joining different data sets to get top claims based on state using **Map Reduce**.
* Converting queries to **Spark SQL** and using parquet file as storage format.
* Developed analytical component using **Scala, Spark and SparkStream**.
* Configured Spark streaming to receive real time data from the **Kafka** and store the stream data to HDFS using **Scala**.
* Written spark programs in Scala and ran spark jobs on **YARN**.
* Designed and Implemented **Solr**Search using the big data pipeline.
* Assembled **Hive and Hbase** with Solr to build a full pipeline for data analysis.
* Written Storm topology to emit data into Cassandra DB.
* Experienced in sync up Solr with **HBase** to compute indexed views for data exploration.
* Implemented **map reduce** programs to perform joins on the Map side using Distributed Cache in Java. Developed Unit test cases using **Junit,** Easy Mock and **MRUnit** testing frameworks.
* Used in depth features of **Tableau** like Data Blending from multiple data sources to attain data analysis.
* Used **Maven** extensively for building MapReduce jar files and deployed it to Amazon Web Services (**AWS**) using EC2 virtual Servers in the cloud.
* Setup Amazon web services (**AWS**) to check whether Hadoop is a feasible solution or not.
* Worked with BI teams in generating the reports and designing **ETL** workflows on **Tableau.**
* Knowledgeable on **Talend** for Data integration purpose.
* Create a complete processing engine, based on Cloudera's distribution, enhanced to performance.
* Experienced in Monitoring Cluster using Cloudera manager.

**Environment:** Hadoop, HDFS, HBase, MapReduce, Java, JDK 1.5, J2EE 1.4, Struts 1.3, Spark, Python, Hive, Pig, Sqoop, Flume, Impala, Oozie, Hue, Solr, Zookeeper, Kafka, GCP, AWS, Cassandra, AVRO Files, SQL, ETL, DWH, Cloudera Manager, Talend, MySQL, Scala, MongoDB.

**Cardinal Health, Dublin, OH July 2016 – Dec 2016**

**Role: Hadoop Developer**

**Description:**

**Cardinal Health** is improving patient care through an optimized supply chain using analytic innovations with Teradata and Hadoop. Employing a unified data strategy has allowed Cardinal Health to see significant business value including a 50%-time savings for end users working with raw data. I was part of the Hadoop development team responsible for improving the costumer experience through Hadoop advantages.

**Responsibilities:**

* Analyzing Hadoop cluster and different Big Data analytic tools including Pig, Hive, HBase and Sqoop.
* Creating multiple MapReduce jobs in Pig and Hive for data cleaning and pre-processing.
* Successfully loading files to Hive and HDFS from Oracle, SQL Server using Sqoop.
* Writing Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Creating Hive tables, loading with data and writing Hive queries.
* Involved in Spark for fast processing of data. Defining job flows.
* Using Hive to analyze the partitioned data and compute various metrics for reporting.
* Moved data from HDFS to Cassandra using MapReduce and BulkOutputFormat class.
* Managing and reviewing theHadoop log files.
* Using Pig as ETL tool to do Transformations, even joins and some pre-aggregations.
* Unit testing and delivered Unit test plans and results documents.
* Exporting data from HDFS environment into RDBMS using Sqoop for report generation and visualization purpose.
* Worked on Oozie workflow engine for job scheduling.

**Environment:** Hadoop, HDFS, MapReduce, Pig, Hive, Sqoop, HBase, Kafka, AWS, Oozie, Zookeeper, Java, Spark, Scala.

**Open Solutions, India Sept 2013 – Sep 2015**

**Role: Java/J2EE Developer**

**Responsibilities:**

* Involved in the process Design, Coding and Testing phases of the software development cycle.
* Designed use-case, sequence and class diagram (**UML**).
* Developed rich web user interfaces using JavaScript (pre-developed library).
* Created modules in **Java and C++, python**.
* Developed **JSP** pages with **Struts framework, Custom tags and JSTL**.
* Developed Servlets, **JSP** pages, **Beans**, **JavaScript** and worked on integration.
* Developed **SOAP/WSDL** interface to exchange usage and Image and terrain information from Geomaps.
* Developed Unit test cases for the classes using **JUnit**.
* Developed **stored procedures** to extract data from **Oracle database**.
* Developed and maintained **Ant Scripts** for the build purposes on testing and production environments.
* Designed and developed user interface components using **AJAX, JQuery, JSON, JSP, JSTL & Custom Tag library**.
* Involved in building and parsing **XML** documents using **SAX** parser.
* Application developed with strict adherence to **J2EE** best practices.

**Environment:** Java, C++, Python, Ajax, JavaScript, Struts, Spring, Hibernate, SQL/PLSQL, Web Services, WSDL, Linux, Unix.