**Dinesh Annapareddy**

**PHONE: (201) 677-5372.**

**Email :** [**dineshklce23@gmail.com**](mailto:dineshklce23@gmail.com)

# PROFESSIONAL SUMMARY

* About 8 years of IT experience in software Design, Development, Analysis, Maintenance, Testing, Support and troubleshooting in **Banking, Retail, manufacturing, brokerage** and **Pharmaceuticals** Industries.
* 5+ years of experience in **installing, configuring, designing, implementing** and **testing** Hadoop ecosystem.
* Around 2 years of experience in data modeling, data analysis, design of **OLTP** and **OLAP** systems.
* Experience in **ETL** (Extract, Transform, and Load) processes using **Informatica Power Center**.
* Experience in designing and developing enterprise applications using **Java**.
* Expert knowledge of major Big Data solutions using **Apache Hadoop, CDH, HDP** and **Teradata** **UDA**.
* Expertise in using various Hadoop ecosystem such as **MapReduce, HDFS, Hive, Pig, impala, R, HBase, Sqoop, Oozie, Zookeeper and flume** for data storage and analysis.
* Contributed towards implementing advanced procedures in **Apache Spark**like text analytics and machine learning applications written in **Java**
* Familiar with **EC2, S3, ELB, cloud watch, SNS, Elastic IP's** and managing security groups, IAM on AWS.
* Have experience with Spark processing Framework such as **Spark** **Streaming** and **Spark Sql**.
* Involved in ingesting data into HDFS using Apache **NIFI.**
* Experience in NoSQL databases **like HBase, Cassandra, Redis, couchDB** and **MangoDB**.
* Hands on working experience with different cloud services like **(AWS), Google** Cloud and **Azure**
* Extensive experience using **Spring Boot Framework** in business layer for Dependency Injection, Security.
* Hands-on experience with message brokers such as **Apache Kafka**, **Oracle AQ** and **Rabbit**
* Experience in Spark Data Frame class capability, transformations and processing semi structured data using **SCALA**.
* Currently working on improving query results for an information retrieval system by using   
  **Apache Solr, Lucene** and Elastic search.
* Extracted data from various data source including OLE DB, **Green plum**, Excel, flat files and XML
* Developed and demonstrated **HIPAA**real-time requests using Spark and storm
* Working on implementing Spark and **Strom** frame work.
* Worked extensively on design, development and deployment of jobs to extract data, filter the data and load them into **Datalake.**
* Developed **Machine learning** application in eDiscovery, content management
* Implemented various **machine learning** techniques like Random Forest, K-Means, Logistic
* Expertise in running query using **Impala** and used BI tools to run ad-hoc queries directly on Hadoop.
* Direct Integration of **Kyvos** with traditional BI tools like Tableau, Excel, and more
* Familiar with Java virtual machine **(JVM)** and multithreaded processing.
* Development experience in **DBMS** like Oracle, MS SQL Server and MYSQL
* Comprehensive knowledge and experience in normalization**/de-normalization**, **data extraction, data cleansing, data manipulation**
* Worked with Kafka and **Strom** for capturing retail-time streaming data.
* Involved in requirements and design phase to implement Streaming **Lambda** Architecture to use real time streaming using Spark and Kafka (version 0.8.2.2).
* Experience in using **CSV, TSV, Sequence files, RC File, JSON, AVRO, PARQUET** and **HAR** file formats.
* Involved in converting Hive/SQL queries into Spark transformations using **Spark RDDs**, **Python** and **Scala**.
* Applied **Spark SQL** and DATAFRAME API in files stored in HDFS
* Improving performance of search results using Spark and **Elastic Search integration**
* Experience in **Behavior Driven Development** and **Test Driven Development**.
* Modeled and analyzed graph data using **Orient DB**.
* Used**SAS** to do data analysis for several projects
* Experience in developing Search engine using **Apache Solr**
* Expertise in developing web applications using **JAVA, Spring MVC**.
* Used **RegEx** and **Avro** **SerDe’**s for serialization and de-serialization packaged with Hive to parse the contents of streamed log unstructured data by implementing Hive custom generic UDF’s.
* Experience with Cluster/Load/State Management using **Zookeeper**.
* COMPILE AND DEBUG HADOOP APPLICATIONS WITH **INTELLIJ**
* Experience in collecting the **log data** from different sources using **Flume** & **Kafka**.
* Excellent working experience in **Jenkins**.
* Performed log analysis on reporting tools like **Splunk** and generated BI reports on **Pentaho**.
* Developed UI using CSS, HTML, JavaScript, **AngularJS**, **JSON**.
* Processing semi-structured **clickstream** data to gain understanding user behavior and target potential new subscribers
* Experience in importing and exporting data using **Sqoop** from HDFS/Hive/HBase to **RDBMS**.
* Created 25+ Linux **Bash scripts** for users, groups, data distribution, capacity planning, and system monitoring.
* Experience in writing custom **UDFs** in **java** for Hive and Pig.
* Developed Pig extracts for migrated **OLAP** reports to Hadoop.
* Responsible for writing Hive Queries for analyzing terabytes of customer data from **HBase.**
* Extensively used **Spark-SQL & Hive Query Language (HQL**) for getting customer insights which will be used for critical decision making by business users.
* Experienced in Installation, Configuration, Upgrade, Migration and Administration of **OBIEE, Oracle BI** Apps
* Worked with the infrastructure teams to configure systems for Hadoop **deployment.**
* Analyzed the SQL scripts and designed the solution to implement using **PySpark**.
* Copy data from **HDFS TO MONGODB** using pig, hive using scripts.
* Managing and scheduling Jobs on a Hadoop cluster using **Oozie, Control-M**.
* Building Fine Grain security in Multi-tenant Hadoop cluster using **Ranger/Knox/Kerberos**
* Recognized as a key subject matter expert for data model and **ETL design**, build and testing initiatives
* Expertise in **Data Science** **- Statistical Analysis** and Machine Learning
* Worked with the Apache **NIFI** flow to perform the conversion of Raw XML data into JSON, AVRO.
* Coded pages with best **Search Engine Optimization** (SEO) techniques and Used JMS to send/receive asynchronous data to other applications
* Implemented ETL processes for warehouse and designed and implemented code for migrating data to **Data lake** using Spark
* Experience with **Korn** scripting
* Worked using AWS Cloud EC2 and data processing using **Redshift** MPP.
* Experience in monitoring and troubleshooting issues with **Linux memory, CPU, OS, storage** and **network**.
* Worked with application teams to install operating system, Hadoop updates, patches, version as required.
* Involved in various phases like Requirements Analysis/Design, Functional Testing **(FUT),** System Integration Testing **(SIT),** Regression testing and Implementation.
* Develop complex sequential and parallel **Data Stage 8.1** jobs.
* Developed, with project managers, a complex project schedule using **Primavera.**
* Setup **Knox** security for the cluster and configured it for web HDFS and HiveServer2
* Responsible for **Web Logic administration**, maintenance and performance tuning
* Experienced in job workflow scheduling and monitoring tools like **Oozie and Zookeeper**.
* Good Exposure in maintaining various version controls like **GIT** and **SVN**.
* Top percentile performer for meeting project deadlines and sales goals while resolving client conflicts.

|  |  |
| --- | --- |
|  | **PROFESSIONAL SKILLS** |
|  |  |
| Big data Ecosystem | Hadoop, MapReduce, HDFS, pig, Hive, Impala, Sqoop, flume, oozie, MangoDB, R, Zookeeper, HBase, Cassandra, Kafka, NIFI, Storm, Solar, Lucene,Kerberos, Spark with Scala, Hadoop Distributions (Cloudera, Hortonworks, datastax, Teradata). |
| Amazon Web Services (AWS) | Elastic Map Reduce (EMR), Amazon EC2, Amazon S3, AWS Code Commit, AWS code Deploy, AWS Code Pipeline, Amazon Cloud Front, AWS Import/Export. |
| Databases | Cassandra, HBase, DB2, Oracle, MS Access, MySQL, MongoDB, PostgreSQL, Microsoft SQL Server, Teradata and IMS. |
| ETL / DWH Tools | Informatica Power Center, Informatica Power Exchange |
| BI / Reporting Tools | Micro Strategy Web […] Hyperion 9.0 Interactive Reporting, Business Objects XIR2, Congos , SSRS, SASD3, NVD3, High charts, Google Chart API, Tableau, Data Meer. |
| Operating system | UNIX, Windows, LINUX. |
| Programming Languages | Java, Scala/ C++, Python, Unix shell. |
| Web Technologies | JSP, HTML, CSS, java script. |
| Web Services | SOAP, Apache CXF/Xfire, Apache Axis, REST Application Servers Apache Tomcat, WEB logic, Web Sphere |
| Testing Tools | JUnit, TestNG, Selenium, NetBeans, Eclipse |
| Methodologies Agile, UML, Waterfall Design Patterns | |

Client Technologies JQUERY, Java Script, AJAX, CSS, HTML 5, XHTML

Other Tools Putty, WinSCP, IntelliJ Data Lake, PyCharam GitHub, SVN, CVS, MS office

**WORK EXPERIENCE**

# Optum , Little rock, AR

# Sr. Hadoop/Spark developer

## **September 2016-Present**

**Project Description:** Optum is healthcare network. In this project we used the core Hadoop ecosystem to build a data lake and retrieve the data of the claims and policy related information. Data is used to keep track of claims and to decommission few legacy systems.

* Extensively used **Spark-SQL & Hive Query Language (HQL**) for getting customer insights which will be used for critical decision making by business users.
* Involved in converting Hive/SQL queries into Spark transformations using **Spark RDDs**, **Python**
* Copy data to **AWS S3** for storage and use COPY command to transfer data to Redshift.
* Developed custom **spark streaming** with **Kafka** for importing data from heterogeneous data sources.
* Migrated 180 tables from **Oracle to Cassandra** using Apache Spark
* Used the **Spark - Cassandra** Connector to load data to and from Cassandra.
* Used **Teradata Connector** for Hadoop (TDCH) for ETL to import e-Scan data (related to Medicare and Medicaid) into HDFS.
* \* Created Presentation catalogs, and Folders in Presentation Layer in**Oracle** BI Administration Tool.
* Implemented many **Impala scripts** and Shell scripts for data validation and data analytics
* Design and implement of **Spark/Solr** solution.
* Developed a data pipeline using Kafka and **Strom** to store data into HDFS.
* implemented different **machine learning** techniques in Scala using Spark machine learning library.
* Worked on data analytics using **R**, Spark and **Python** using machine learning libraries.
* **Data ingestion** from various data sources to HBase and Hive using Sqoop and Spark.
* Worked on various **HBase** compression techniques such as **Sanppy, Encoding FAST\_DIFF**.
* Developed Spark scripts by using **Python** as per the requirement.
* Used **SAS**Data integration studio for registering the tables and creating the libraries pointing to Teradata database
* Worked with **Apache Avro** for encoding and **data serialization**.
* Apache **Oozie** Workflow Scheduler for Hadoop.
* Developed and deployed Apache **NIFI** flows across various environments, optimized **NIFI** data flows and written QA scripts in python for tracking missing files.
* Responsible for building **pipelines** to get data from **Azure, AWS to Spark**.
* Implemented **security** on Cloudera Hadoop
* Experienced in providing security to Hadoop cluster with **Kerberos** and integration
* Prepared Data Dictionaries for the **Tableau** Data source
* Prepare Monthly and quarterly reports using **Kyvos** **visualization tools.**
* Experience working in an **Agile** environment with fast paced development and sprints to keep track of the progress of the project.
* Incorporated custom logging mechanism for tracing errors, resolving all issues and bugs before deploying the application in the **WebSphere** Server.
* Creating and Building Datasets, Viz boards **(Dashboards)** and Lens on **Platfora**

## **AAA Insurance , Emeryville, CA**

# Sr. Hadoop developer

## **January 2016-August 2016**

## **Project Description**: AAA Insurance group is one of the biggest insurance groups. It processes 3 million claims daily we receive the insurance claim data from auto adjudication system in X12 format, which is fixed layout format. We are loading the data into HDFS and had written multiple map reduce jobs to convert the X12 format into CSV format and loading the data into Hive after creating the tables. Hive join queries are using to fetch information from multiple tables. The queries output is collected into temporary tables to perform more complex joins.

* Developed **Spark** scripts by using **Python** commands as per the requirement.
* Developed multiple MapReduce jobs in Java for **data cleaning** and preprocessing
* Used with NoSQL technology **(Amazon Dynodb)** to gather and track event-based metrics.
* Good Experience in writing UDFs, **Pig scripts** and **Hive Queries** for processing and analyzing large volumes of data.
* Imported data from Oracle database to HDFS using Unix based **File Watcher tool**.
* Worked on Data Serialization formats for converting Complex objects into sequence bits by using **AVRO, PARQUET, JSON, CSV** formats.
* Implemented **Kerberos** Security Authentication protocol for existing cluster.
* Writing Sqoop scripts to import data historical data and used CDC to capture Transactional data to scalable big data store called a **DATALAKE** Used R to create analysis programs using linear regression, pattern recognition, and machine learning methods.
* Implemented **NIFI** flow topologies to perform cleansing operations before moving data into HDFS.
* Worked on implementing Spark and **Strom** frame work.
* Used **SAS/STAT** to create different statistical reports as required
* Maintenance of all the services in Hadoop ecosystem using **ZOOKEPER**.
* Designed **Hive Schemas** on Procurement Order and Production Order quantity from fiscal year sales forecast data using partitioned and bucketed columns/data and run queries using **partition/bucketed** Columns.
* Experience in using Impala and **Hive Server2** using the **Squirrel Client**
* Used Hive to analyze data ingested into HBase by using **Hive-HBase** integration and compute various metrics for reporting on the dashboard
* Detected and deployed hotfixes to resolve over **300 production bugs**, incompatibilities, and other issues
* Created **HBase** tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, **NoSQL** and a variety of portfolios.
* Used **OOZIE** Operational Services for batch processing and scheduling workflows dynamically.
* creation of **MongoDB** based Hive table.
* Analyzed & designed the business rules for data cleansing that are required by the staging & **OLAP** and **OLTP** databases.
* Setting up NIS, NFS, DNS and configured **Mail server**.
* Worked on the **proof-of-concept** for Apache Hadoop framework initiation.
* Worked using AWS Cloud EC2 and data processing using **Redshift** MPP
* Streamlined teams ingest/egress process and flow using Agile tool **(JIRA)**
* Amazon **Redshift to Tableau** Integration for reporting & visualizations

## **Next Education, Hyderabad, INDIA**

## **Hadoop developer**

**January 2014-July 2015**

**Project Description:** Next Education Data Engineering team had developed and supported more than 20 software applications for its improvising its business. To handle high volume data from various frontline systems, Databases like Oracle and offload its ETL process from Informatica, built 150 node Hadoop Cluster. Some of the data is directly ingested from front line systems into the cluster using Flume. This cluster is the source system for various ETL reports and analytics applications.

* Experience in Hadoop administration
* ion activities such as installation and configuration of clusters using **Cloudera** and **Hartonworks**
* Developed **SPARK** CODE using **Python** and **Spark-SQL/Streaming** for faster testing and processing of data.
* Wrote **scala scripts** as required for spark engine.
* Working on various Use Cases on **data ingestion** from various resources into Hadoop Cluster.
* Data migrating from various sources using Unix, Sqoop, Oozie, Hive 0.13 with **CDH 5.3.2**
* Use of **Unix shell script** to bring the data from various sources into Hadoop Cluster
* Involved in moving all log files generated from various sources to HDFS for further processing through **Flume**.
* Data is processed using **java** based Map reduce jobs, Hive and Pig Queries based on the complexity
* Created **HiveQL** query, Map Reduce Job for reporting purpose
* **Merge** multiple ingested files in Hadoop using pig. Debug **pig scripts** .
* Integrated the hive warehouse with **HBase** and **Couch DB.**
* Involved in running query using **Impala** and used BI tools to run ad-hoc queries directly on   
  Hadoop.
* Exported data from **MySQL, oracle database, Teradata** to HDFS using Sqoop
* Extracted feeds form **social media** sites such as Facebook, Twitter using Java scripts
* Worked in **Apache Tomcat** for deploying and testing the application.
* Integrate Hadoop with DB2 to support business intelligence reporting with **Cognos**
* Cost Savings on Hadoop Clusters in **Amazon EMR** (Strategies like Reserved instance, spot instances and proactive unwanted cluster termination.)
* Led implementation on a **150 node** Hadoop cluster for a monthly data in-flow of approx. 2 TB.
* Used both **Sqoop 1** and **Sqoop 2** to materialize and transfer datasets between Hive and our Data Warehouse.
* Used **Maven, Eclipse** and **Ant** to build the application

**Ford Motor Company, Chennai, India**

**Hadoop developer**

**October 2012- December 2013**

**Project Description:** All the customer and accounting information are stored in RDBMS and collecting large amount of log data from distributed sources. volume of data is huge and growing in RDBMS and it has high maintenance cost, Wells Fargo started to move the data to HDFS as a part of RDBMS to Hadoop migration. Hadoop helps to implement Business rules, predictive analytics on risks, compliance and marketing.

* Maintenance of Hadoop Clusters using **HDP** and **CDH3.**
* upgrades of Cloudera Hadoop distribution (**CDH3 to CDH4**).
* Created J**ava scripts** based upon the luigi framework that do things like reach out to FTP locations, stream multiple files into one and store the resulting file on **HDFS.**
* Developed Hadoop/Hive based features in **Java.**
* Exported summary tables into **MySQL** for reporting purposes using Sqoop
* Extracted data from **oracle database** to HDFS using Sqoop.
* Handled importing of data from machine logs using **Flume.**
* Handled incremental data loads from **Teradata** into HDFS.
* Performed **Map Reduce** programs on **log data** to transform into structured way to find user location, age group, spending time.
* Handled importing of data from various data sources, performed transformations using **hive**. (External tables, partitioning).
* Import & Export of data from one server to other servers using tools like Data Transformation Services **(DTS)**.
* Performed analysis of data using **pig scripting**.
* Converted **SQL scripts** into hive to get better performance.
* Developed a **web crawler** code to obtain the raw data of product review and performed **data cleansing** in Python.
* Implemented Custom File loader for Pig so that we can query directly on the large Data files suchasbuildlogs   
  Improved the Performance by tuning of HIVE and map reduce.
* Created connection through **JDBC.**
* Created and maintained Technical documentation for launching HADOOP Clusters and for executing **Hive queries, Pig Scripts,** **Sqoop** job.
* Prepare **multi-cluster** test harness to exercise the system for performance and failover.
* Involved in design and development of **UI** component, using frameworks Angular JS.

**Karvy Financial, Hyderabad, INDIA**

**ETL/Informatica developer**

## **May2010-September2012**

**Project Description**: The goal of the project was to design and develop a broker/dealer data mart which would be used to calculate the commissions to be paid to reps based on their sales. It would also help the agencies analyze the sales of the different products. Source feeds were from Oracle, DB2 and flat files and the target tables resided in Oracle.

* Translating the functional requirements into technical mapping specifications and **leveraged Informatica Power Center 8.6.1** to extract the data from AS/400 DB2 files, Oracle tables and flat files to populate billing, service order and plant interface files, which are then loaded into the DPI Billing system.
* FTP scripts and Extraction of Unstructured Data from various karvy Data formats using **Perl/Shell Scripts**. XML/flat file formats also extracted successfully.
* Exract /Transform/Load (ETL) processes in**Pentaho** Data Integration (kettle).
* Design and implement the application for consuming the source (Flat file, Web Service, MQ- Series real time) feed data from upstream.
* Design, develop, and deploy ETL processes and performance tune **ETL programs/script.**
* Design and development of complex **ETL mappings** making use of Connected/Unconnected Lookups, Normalizer, Stored Procedures, SQL & Custom Transformations.

* Implementation of successful strategies in loading feeds to the Frontier Data Warehouse.
* Optimization/performance tuning techniques implemented to identify **bottlenecks- Query** tuning (Explain Plan, SQL Trace), table partitioning, memory tuning and cache management.
* Development of PL/SQL Stored Procedures to implement complex business logic and integrate them with the ETL process as pre/post Stored Procedures.
* SQL Loader scripts for bulk data loads to Oracle tables and extracting Oracle dumps from karvy.

## **Aircel Telecommunications, Chennai, India**

# ETL/Informatica developer

## **June2008-April2010**

## 

## **Project Description:** This project deals about the enterprise payment system in Aircel to generate the bill due amount based on the payment received for the policies from the customers directly/indirectly through payment channels, third party collection centers and real-time/online services. ETL batch service will receive the bill due feed from several billing systems daily and ETL process will cleanse and validate the data before passing through the centralized payment central hub database. After the required validation and applying the business logic .

## Created Technical Design Document / BRD along with Data Discovery sheet .

## Closely worked with SME's Application owners and end users to get details on their apps and requirement s .

## Extensively worked on Data Masking transformation to mask data in **DEV/QA/SIT** environments.

## Worked on ILM to create and define rules for **data masking**.

## Worked on Performance Tuning to optimize the Session performance by utilizing, Partitioning, Push down.

## optimization, Index Cache, Data Cache and incremental aggregation.

## Worked on different databases such as **Oracle, DB2** and **Teradata .**

## Implemented various Transformations like Joiner, Aggregate, Expression, Lookup, Filter, Update Strategy etc .

## Used **Power Exchange** for Mainframe VSAM file .

## Used Target Load Plan to execute the mappings sequentially in a pipeline structure .

## Migrating and deploying the ETL code in **QA** and **UAT** environments.

## Used Constraint Based Loading, partitioning, performance tuning on existing mapping .

## Worded with Change Management Team to roll-out final solution .

Created **UNIX script** to run batch jobs per business requirements .