**NAGESWARA,YOGESH**

Sr Hadoop Spark Developer

E-mail: [yogeshnages46@gmail.com](mailto:yogeshnages46@gmail.com);

Phone: (570)-435-1649

**Professional Summary:**

* Professional experience of **8+ years** in IT which includes 4 years of comprehensive experience in working with **Apache Hadoop Ecosystem-components, Spark streaming** and **Amazon Web services(AWS)**.
* Over **5+ Years** of development experience in Big Data Hadoop Ecosystem components and related tools with data ingestion, importing, exporting, storage, querying, pre-processing and analysing of big data.
* Good Working Expertise on handling Terabytes of structured and unstructured data onhuge Cluster environment.
* Experience in using **SDLC** methodologies like **Waterfall, Agile Scrum, TDD** for design and development.
* Expertise in implementing **Spark** modules and tuning its performance.
* Experienced in performance tuning of Spark applications using various **resource allocation** techniques and**transformations,** reducing the **Shuffles** and increasing the **Data Locality configurations**.
* Expertise in Kerberos Security Implementation and securing the cluster.
* Expertise in creating **Hive Internal/External Tables/Views** using shared Meta store, writing scripts in **Havel** and experience in data transformation & file processing, building analytics using **Pig Latin** Scripts.
* Expertise in writing custom UDFs in **Pig & Hive** Core Functionality.
* Hands on experience with AWS components like **EC2, EMR** and **S3**.
* Developed, deployed and supported several **Map Reduce** applications in Java to handle different types of data.
* Worked with various compression techniques like Avro, Snappy, LZO.
* Hands on experience dealing with **AVRO** and **Parquet** file format, following best Practices and improving the performance using **Partitioning, Bucketing, Map side-joins** and creating **Indexes**.
* Expert in implementing advanced procedures like text analytics, processing and implementing **streaming APIs** using the in-memory computing capabilities like **Apache Spark** written in **Scala, Python** and **Scala.**
* Designed and configured Azure Virtual Networks (VNets), subnets, Azure network settings, DHCP address blocks, DNS settings, security policies and routing.
* Experience in Data Load Management, importing and exporting data from HDFS to Relational and non- Relational Database Systems using **Sqoop, Flume** and **Apache Nifi** by efficient column mappings and maintaining the uniformity**.**
* Exported data to various Databases like **Teradata** (Sales Data Warehouse), SQL-Server, Cassandra using **Sqoop.**
* Experienced in creating shell scripts to push data loads from various sources from the **edge nodes** onto the HDFS.
* Experienced in performing code reviews, involved closely in smoke testing sessions, retrospective sessions.
* Experience in scheduling and monitoring jobs using **Oozie** and **Crontab.**
* Experienced in Microsoft Business Intelligence tools, developing **SSIS (Integration Service), SSAS (Analysis Service)and SSRS (Reporting Service)**, building **Key Performance Indicators** and **OLAP** cubes.
* Have hands on experience in creating cubes for various reports for end clients, Configured Data Source and **Data Source Views, Dimensions, Cubes, Measures, Partitions, KPI’s and MDX Queries**.
* Have good exposure with the star, snow flake schema, data modelling and work with different data warehouse projects.
* Hands on working with the reporting tool Tableau, creating dashboards attractive dashboards and worksheets.
* Extensive work experience with **Java/J2EE** technologies such **as Servlets, JSP, EJB, JDBC, JSF, Struts, spring,SOA, AJAX, XML/XSL, Web Services (REST, SOAP), UML, Design Patterns** and **XML Schemas.**
* Strong experience in design and development of relational database concepts with multiple RDBMS databasesincluding **Oracle10g, MySQL, MS SQL Server & PL/SQL.**
* Experience **in JAVA, J2EE, WEB SERVICES, SOAP, HTML** and **XML** related technologies.
* Have closely worked with the technical teams, business teams and product owners.
* Strong analytical and problem-solving skills and ability to follow through with projects from inception to completion.
* Ability to work effectively in cross-functional team environments, excellent communication and interpersonal skills.

**Tools and Technologies:**

|  |  |
| --- | --- |
| **Hadoop/BigData Technologies** | **HDFS, Map Reduce, Sqoop, Flume, Pig, Hive, Oozie, Impala, Zookeeper, Ambary, Storm, Spark and Kafka** |
| **No SQL Database** | **HBase, Cassandra, MongoDB** |
| **Monitoring and Reporting** | **Tableau, Custom Shell Scripts** |
| **Hadoop Distribution** | **Horton Works, Cloudera, MapR** |
| **Build Tools** | **Maven, SQL Developer** |
| **Programming and Scripting** | **Java, C, C++, C# on .net, JavaScript, Shell Scripting, Python, Scala, Pig Latin, HiveQL** |
| **Java Technologies** | **Servlets, JavaBeans, JDBC, Spring, Hibernate, SOAP/REST services** |
| **Databases** | **Oracle, MY SQL, MS SQL server, Vertica, Teradata** |
| **Analytics Tools** | **Tableau, Microsoft SSIS, SSAS and SSRS** |
| **Web Dev. Technologies** | **HTML, XML, JSON, CSS, JQUERY, JavaScript** |
| **IDE Dev. Tools** | **Eclipse 3.5, Net Beans, My Eclipse, Oracle, JDeveloper 10.1.3, SOAP UI, Ant, Maven, RAD** |
| **Operating Systems** | **Linux, Unix, Windows 8, Windows 7, Windows Server 2008/2003** |
| **Hadoop/Big Data Technologies** | **HDFS, Map Reduce, Sqoop, Flume, Pig, Hive, Oozie, Impala, Zookeeper, Ambary, Storm, Spark and Kafka, Apache Nifi** |
| **Network protocols** | **TCP/IP, UDP, HTTP, DNS, DHCP** |

**Professional Experience:**

**Apple Inc- Sunnyvale, CA October 2016–Till date**

**Sr Hadoop Spark Developer**

**Description:** The company is committed to bringing the best personal computing and music experience to students, educators, creative professionals, businesses, government agencies, and consumers through its innovative hardware, software, peripherals, services, and Internet offerings. The project mainly concentrated on building analytics on top of device usage information collected from various IOS applications and devices.The company’s business strategy leverages through the design and development of its own operating system, hardware, and many software applications and technologies, to bring to its customers new products and solutions with superior ease-of- use, seamless integration, and innovative industrial design.

**Responsibilities:**

* Worked on different file formats like Sequence, **XML, JSON** files and Map files using Map ReducePrograms.
* Took important decisions of how much the **poll time** should be for the stream processing, what type of Hadoopstack component to use for better performance.
* Proposed and implemented a solution for their long-time issue of ordering the data in **Kafka** queues.
* Designed and implemented an **ETL** framework with the help of sqoop, pig and hive to be able to automate the process of frequently bringing in data from the source and make it available for consumption.
* Worked on importing and exporting data into **HDFS** and **Hive** using **Sqoop,** built analytics on **Hive** tables using **HiveContext**.
* Developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with Pig.
* Load and transform large sets of semi-structured and unstructured data on **HBase** and **Hive**.
* ImplementedMap-Reduceprogrammingmodelwith**XML,JSON,CSV**fileformats.Made use of **SERDE** jars to load json and xml format dataon to Hive tables coming from Kafka queues.
* Designed and configured Azure Virtual Networks (VNets), subnets, Azure network settings, DHCP address blocks, DNS settings, security policies and routing.
* Implemented UDFs for **Hive** extending **Generic UDF, UDTF** and **UDAF** base classes to change the time zones, implement logic actions and extract required parameters according to the business specification.
* Extensive working knowledge of **Partitioning**, **UDFs**, **Performance tuning, Compression**-related properties on Hivetables.
* Developed the UNIX shell scripts for creating the reports from Hive data.
* Implemented **Spark** scripts in Python to perform extraction of required data from the data sets and storing it on HDFS.
* Developed spark scripts and python functions that involve performing transformations and actions on data sets.
* Configuring**Spark**Streaming in Pythontoreceiverealtimedatafromthe**Kafka**andstoreit onto**HDFS.**
* Experienced in building analytics on top of spark using machine learning**Spark.ml**.
* Involved in optimizing the Hive queries using **Map-side join, Partitioning, Bucketing and Indexing**.
* Involved in tuning the Spark modules with various memory and resource allocation parameters, setting right Batch Interval time and varying the number of executorsto meet the increasing load overtime.
* Continuously monitored and managed the Hadoop cluster using Cloudera Manager.
* Used Hue for UI based PIG script execution, Tidalscheduling and creating tables in Hive.
* Created **Pig Latin** scripts to sort, group, join and filter the enterprise wisedata.
* Involved in planning process of iterations under the Agile Scrum methodology.
* Extensive knowledge of working on Apache **NiFi,** used and configured of different processors to pre-process, make the incoming data uniform and format according to the requirement.
* Implemented unit testing in Java for pig and hive applications.
* Hands on experience in **AWS** Cloud in various AWS services such as **Red shift cluster, Route 53**domain configuration.
* Extensively used UNIX for shell Scripting and pulling the Logs from the Server and monitor it.
* Worked with the Data Science team to gather requirements for various data miningprojects.

**Environment: Cloudera CDH 5.1.4, Apache Hadoop 2.3.0 (Yarn), Spark 2.1.0,Spark.ml, Flume 1.7.0, Eclipse, Map Reduce, Hive 1.2.2, Pig Latin 0.17.0, Java, SQL, Sqoop 1.4.6, Centos, Azure, Zookeeper 3.5.0 and NOSQL database, Apache Nifi, AWS, S3, EMR, Red Shift Cluster.**

# Blue Shields of Arizona- Phoenix, AZ July 2015 - September 2016

# Hadoop Developer

# Description:Blue shields of Arizona, a Fortune 500 company, is a diversified, healthcare enterprisethat provides services to government-sponsored healthcare programs, focusing on under-insuredand uninsured individuals. Many receive benefits provided under Medicaid.The project aims at importing customer details from various heterogeneous sources onto HDFS. And developing a web application that facilitates the customers to easily access their plans, claims and details of their coverage as well as their history of claims and bills.

# Responsibilities:

* + Worked on extracting data from **Oracle** database and load to **Hive**database.
  + Used Spark-Streaming APIs to perform necessary transformations and actions on the fly from **Kafka**queues in real time and persist on **Cassandra**using the required connectors and drivers.
  + Integrated Kafka, Spark and Cassandra for streamline analytics for creating a predictive model.
  + Developed Scala scripts, UDFs using both Data frames in Spark for Data Aggregation, queries and writing data back into **OLTP** system through Sqoop.
  + Worked on modifying and executing the UNIX shell scripts files for processing data and loading to HDFS.
  + Worked extensively on optimizing transformations for betterperformance.
  + Was involved in carrying out the important design decisions in creating **UDFs, partitioning** the data in hive tables at two different levels based on the related columns for efficient retrieval and processing of queries.
  + Create Cache Memory on Windows Azure to improve the performance of data transfer between SQL Azure and WCF services.
  + Tweaked lot of options to get performance boost like trying it out with different executer count and memoryoptions.
  + My team was also involved in maintenance, adding the feature of stable time zones across all records in the database.
  + Uploaded and processed more than 30 terabytes of data from various structured and unstructured, heterogeneous sources into HDFS file systemusing **Sqoop** and **Flume**enforcing and maintaining the uniformity across all the tables.
  + Developed complex transformations using **HiveQL** to build aggregate/summarytables.
  + Used **Solr/Lucene** for indexing and querying the **JSON** formatted data.
  + Implemented unit tests for pig and hive applications.
  + Developed UDF's in Python to implement functions according to the specifications.
  + Developed Spark scripts, configured according to business logic, good knowledge of actionsavailable.
  + Well versed with the **HL7** international standards as the data was organized according to this format.
  + Formatted and built analytics on top of the data sets that were complied with **HL7** standards.
  + Created UDFs in **scala**, java for formatting and applying transformations on the information in HL7 versions.
  + Analyze the JSON data using hive SerDe API to deserialize and convert into readableformat.
  + Involved in increasing and optimizing the performance of the application using **Partitioning and Bucketing** on Hive tables, developing efficient queries by using **Map-side joins** and **Indexes**.
  + Worked with downstream team in generating the reports on Tableau.
  + Conducted code reviews to ensure systems operations.

**Environment: CDH 5.1.x, Hadoop 2.2.0, HDFS, Map Reduce, Sqoop, Flume, Hive 2.0.x, SQL Server, Azure, TOAD, Oracle, Scala 2.9.1,Solr/Lucene, PL/SQL, Eclipse, JAVA, Shell scripting, Vertica, Unix, Cassandra, HL7 standard.**

**State Farm Insurance- Bloomington, ILAugust 2014 – June 2015**

**Hadoop Developer**

**Description:** State Farm is a large group of insurance and financial services companies in the United States. The group's main business is State Farm Mutual Automobile Insurance Company, a mutual insurance firm that also owns the other State Farm companies. The project mainly deals with designing and implementing a system to receive the data sets, pre-processed and store it in a formatted order from their legacy Teradata database. Also, it deals with building a complete ETL framework by implementing customized Teradata JDBC connectors and process the retrieved information for the business intelligence and Tableau team to make efficient decisions for increasing the productivity and revenue for the organization.

# Responsibilities:

* + Involved in architecture design, development and implementation of Hadoop deployment, backup and recovery systems.
  + Developed MapReduce programs in Python using **Hadoop streaming API** to parse the raw data, populate staging tables and store therefined data in partitionedHIVE tables.
  + Enabled speedy reviews and first mover advantages by using **Oozie** to automate data loading into the Hadoop DistributedFile System and **Pig** to pre-process thedata.
  + Convertedapplications which was on map-reduce architecture to Spark using **Python API** which performed thebusinesslogic.
  + Involved in creating Hive tables, loading with data, writing hive queries that will run internally in map reduce way.
  + Imported **Teradata** datasets onto the **HIVE** platform using **Teradata JDBC connectors**.
  + Was involved in writing **FastLoad** and **MultiLoad** scripts to load the tables
  + Worked with different types of **Indexes and Collect Statistics** in Teradata and improving of execution strategy.
  + Worked with the**SQL assistant and BTEQ**toingest and execute queries, stored procedures and update the tables.
  + Worked in extracting **XML** type files using **XPath** and storing it onto to Hive tables.
  + Developed multiple **Kafka Producers and Consumers** as per the software requirement specifications.
  + Involved in designing the tables in Teradata while importing the data.
  + Developed the UNIX shell scripts for creating the reports from Hive data.
* Experienced in managing and reviewing the Hadoop log files.
* Developed Hive jobs to parse the logs, structure them in tabular format to facilitate effective querying on the log data.
* Extensively used **UNIX** for shell Scripting and pulling the Logs from the Server.
* Worked on different file formats like Sequence files, XML files and Map files using Map Reduce Programs.
  + Worked with **Avro Data Serialization** system to work with **JSON** data formats.
* Implemented the workflows using **Apache Oozie** framework to automate tasks.
  + Completed testing of integration and tracked and solved defects.
  + Used **AWS** services like **EC2** and **S3** for small data sets.
  + Involved in loading data from UNIX file system to HDFS

**Environment: Hadoop Hortonworks 1.3, Hive, Pig, HBASE, Sqoop and Flume, Oozie, AWS, S3, EC2, EMR Spring, Kafka,SQL Assistant, Python,UNIX, Teradata**

**DPSG Inc- Plano, TXMarch 2012 – July 2013**

# Jr. Hadoop Developer

# Description: Dr Pepper Snapple Group, Inc. is an American soft drink and beverage company, based in Plano, Texas. The project was to achieve the task of importing all the historical, legacy and recent data from various database sources by stabilizing the overall data and maintaining the uniformity and store it onto HDFS for frequent retrieval.Also, formatting the data using various techniquesand storing onto HDFS which contained the customer details, production data, sales databased on different flavors and controlling the database’s performance as the load increased overtime were the main aspects.

# Responsibilities:

* Supported and monitored **MapReduce** Programs running on thecluster.
* Evaluated business requirements and prepared detailed specifications that follow project guidelines to developprograms.
* Configured the **Hadoop cluster** with Name-node, Job-tracker, Task-trackers on slavenodes and formatted**HDFS**.
* Used **Oozie** workflow engine to schedule and execute multiple **Hive,Pig**and Spark jobs by passing arguments.
* Involved in creating Hive Tables, loading data and writing Hive queries to invoke and run Map Reduce jobs in thebackend.
* Designed and implemented **Incremental** Imports into Hive tables using Sqoop and cleaning the staging tables and files.
* Involved in collecting, pre-processing,aggregating and moving data from servers to HDFS using **Apache Flume**.
* Developed multiple MapReduce jobs in python for data cleaning andpreprocessing.
* Exported the result set from **Hive** to **MySQL** using **Sqoop** after processing thedata.
* Analyzed the data by performing **Hive** queries and running **Pig** scripts to study customerbehavior.
* Have hands on experience working on **Sequence files**, **AVRO**, **HAR** file formats andcompression.
* Used **Partitioning** and **Bucketing** in Hive tablesto increase the performance and parallelism.
* Experience in writing **MapReduce** programs in pythonto cleanse Structured and unstructureddata.
* Wrote **Pig** Scripts to perform **ETL** procedures on the data in**HDFS**.
* Implemented Hive and Pig scripts to analyze large data sets.
* Loaded and transformed large sets of structured, semi structured and unstructured data onto **HBase**column families.
* Created **HBase** tables to store data coming from differentportfolios and created Bloom filters on column families for efficiency.
* Worked on improving the performance of existing **Pig** and **Hive**Queries by optimization.
* Analyzed the partitioned and bucketed data and compute various metrics forreporting.
* Deployed and worked with **Apache Solr** search engine server to help speed up the search of the sales and production data.
* Written Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the data pertaining to different states across northern-USA.
* Developed workflow in **Oozie** to automate the tasks of loading the data into HDFS and pre-processing with **Pig Latin** scriptsand store these intermediate results for further analytics.
* Migrated **ETL** jobs to **Pig** scripts to perform the **Transformations**, **joins** and pre-aggregations before storing data onto HDFS.
* Involved in loading data from **RDBMS** and web logs into **HDFS** using **Sqoop** and**Flume**.
* Worked on loading the data from **MySQL** to **HBase** where necessary using**Sqoop**
* Co-ordinated with offshore and onsite team to understand the requirements to propagate and prepare designdocuments from the requirements specification and architectural designs.
* Worked with application teams to install Operating Systems, Hadoop updates, patches, and version upgrades as required.

**Environment:Hadoop1x,Hive,Pig,HBASE,SqoopandFlume,Spring,jQuery,Java,J2EE,HTML,JavaScript,Hibernate, PL/SQL, Windows NT, UNIX Shell Scripting, Putty and Eclipse.**

**American National Bank - Omaha, NE October 2010 – February 2012**

**Java developer**

**Description:** American National Bank provides banking products and services to individuals, families, and businesses in Nebraska and Iowa. It offers business banking services that include business loans, corporate banking, commercial real estate, small business, agricultural lending, specialty finance, and treasury services. The aim was to support the design and development team, help them to build a web application for range of customers to access many services efficiently,design and implement interactive front-end, resolve all the production issues and maintain the consistency across the backend database effectively.

# Responsibilities:

* + Implemented **Microsoft Visio** and **Rational Rose** for designing the Use Case Diagrams, Class models, Sequence diagrams, and Activity diagrams for **SDLC** process of theapplication.
  + Deployed GUI pages by using **JSP, JSTL, HTML, DHTML, XHTML, CSS, JavaScript,AJAX**
  + Configured the project on WebSphere 6.1 applicationservers
  + Implemented the online application by using **Core Java, Jdbc, JSP, Servlets and EJB 1.1, WebServices, SOAP,WSDL**
  + Used Log4J logging framework to write Log messages with various levels.
  + Involved in fixing bugs and minor enhancements for the front-end modules.
  + Performed live demos of functional and technical reviews.
  + Maintenance in the testing team for **System testing/Integration/UAT**.
  + Guaranteeing quality in the deliverables to the product owners and business team.
  + Conducted Design reviews and Technical reviews with other project stakeholders.
  + Implemented Action Classes and Server-side validations for account activity, registration and Transaction’s history.
  + Designed user-friendly GUI interface and Web pages using **HTML, CSS, Struts, JSP**.
  + Involved in writing Client-Side Scripts using Java Scripts and Server-Side scripts using **Java Beans**.

**Environment: JDK 1.5, JSP, WebSphere, JDBC, EJB2.0, XML, DOM, SAX, XSLT, CSS, HTML, JNDI, WEB SERVICES, WSDL, SOAP, RAD, PL/SQL, JavaScript, DHTML, XHTML, JavaMail,PL/SQL DEVELOPER, TOAD, POI REPORTS, WINDOWS XP, RED HAT LINUX**

**Pixton Technologies, Bangalore, IndiaSeptember 2009 – August 2010**

****

**Web Developer**

**Description**:Pixton Technologies is a dedicated team of technical professionals having expertise in the field of Media, 2D & 3D Animation, Interiors & Exteriors, Designing’s, Web designing & Development, Advertising, Event Management, Film Making, Electronics, Electricals, Mobile App Development. Main aim of the project to build web application and services for a health app which tracks the details and the conditions of the registered patients and provide a seamless connection between the doctors and patients. Also, deals with customer issues to propose solutions along with configuration and support for enterprise software applications.

# Responsibilities:

* Involved in various stages of Enhancements in the Application by doing the required analysis, development, andtesting.
* Prepared the High and Low-level design document and Generating DigitalSignature
* For analysis and design of application created **Use Cases, Class and SequenceDiagrams.**
* For the registration and validation of the enrolling customer developed logic andcode.
* Developed web-based user interfaces using struts framework.
* Coded and deployed JDBC connectivity in the Servlets to access the Oracle database tables on Tomcat web-server.
* Handled Client-side Validations used **JavaScript**
* Involved in integration of various Struts actions in theframework.
* Used Validation Framework for Server-sideValidations
* Created test cases for the Unit and Integrationtesting.
* Front-end was integrated with Oracle database using **JDBC** API through **JDBC-ODBC Bridge driver**at serverside.

**Environment: Java Servlets, JSP, Java Script, XML, HTML, UML, Apache Tomcat, JDBC, Oracle, SQL, Log4j**