** **

**Big Data Hadoop Engineer /Solutions Architect**

**Name: Anil Kumar U Email: anil.veeras@gmail.com**

**Phone: 754-317-2260**

**SKILLS INVENTORY**

* Around 9 years of professional IT experience which includes 6+ years of proven experience in Big Data Technologies / Framework like Hadoop, Apache Spark, Big Data Application Development And Administration, MongoDB Development and Administration, Analytics and Visualization
* Extensive experience and knowledge of processing BigData using Hadoop ecosystem components HDFS, Map Reduce, Hive, Pig, Sqoop, Oozie, Flume, Hue and NOSQL Data bases as MongoDB
* Implemented multiple projects in next generation big data technology like Spark, Kafka and Elastic Search.
* Designing and deploying scalable, highly available, and fault tolerant systems on AWS
* Designing cloud based optimized and cost effective architectures and systems
* Expert design, built massively scalable Hadoop based distributed data processing systems & applications
* Expertise in data Integration strategies and building next generation DW/BI Infrastructure supporting peta byte scale data in operational/real time reporting and analytic platforms.
* Driving Innovation in creating next generation Enterprise solutions, delivering strategies from infrastructure planning to effective data acquisition, transformation to innovative insights/analytics.
* Domain experience in eCommerce personalization, decision engine, and recommendation Engines.
* Provide oversight of projects ensuring architectural alignment of a project through a full lifecycle: Initial kickoff through requirements analysis, design and implementation to production and launch.
* Driving eCommerce business with advanced analytics, data science, and Unlocking Business Insights.
* Setup Hadoop cluster as a Symantec Consumer centralized data collection platform runs in batch mode.
* Lead efforts to define the overall technology strategy and ensure the realization of goals such as reuse, infrastructure simplification, scalability, cost management, flexibility & HA Strategies
* Well versed in installing, upgrading and maintaining Cloudera (CDH5) distributions for Hadoop.
* Experience in Cloud Hadoop frameworks like AWS (EMR). Hands-on experience with Hadoop applications (such as administration, configuration management, monitoring, debugging, and performance tuning).
* Experience in installation, configuration, management and development of big data solutions using open source Apache Hadoop, CLOUDERA (CDH4, CDH5) & **MapR(M3,M5,M7)**distributions,HDP 2.5
* Extensive Knowledge and working experience with **Hive**
* **Knowledge** On Pre sales related to Insurance and Banking domain in Big Data Hadoop
* Extensive experience in designing MongoDB multi-sharded cluster and monitoring with MMS.
* Experience in developing application, automated scripts leveraging **MongoDB**.
* Expertise knowledge of modifications required in static IP (interfaces), hosts, setting up password-less SSH and Hadoop configuration for **Cluster setup and maintenance**.
* Experience in Design & Setup of Security Architecture involving LDAPS, AD, Kerberos , Knox and Ranger
* Performance tuning of various components in HDP to get the customer jobs to execute within the SLA
* Limited working Experience with AWS
* Have hands on experience in writing **Map Reduce** jobs in Java, Hive, and Pig
* Having good knowledge on **Apache Spark and Spark Streaming in MapR**
* Worked on **Spark Streaming** for **Social Media Analytics(Twitter)**
* Had good Experience in building **Spark Cluster** with several nodes
* Having experience on creating databases, tables and views in **HIVEQL, IMPALA and PIG LATIN**.
* Having knowledge on using **Tableau, QlikView**Reporting Tools
* Knowledge on NoSQL Databases like **MongoDb, Cassandra**, **Hbase**
* Worked on Predictive Algorithms using **R-Scripts**
* Processing of online streaming data according to Client requirements using **Flume**
* Expertise in interacting with business users for requirement analysis and to define business and functional specifications
* Excellent analytical, problem solving, communication and interpersonal skills with ability to interact with individuals at all levels and can work as a part of a team as well as independently
* Strong understanding backed by real time implementations of NoSql databases like Cassandra and MongoDB.
* Strong knowledge on Hadoop HDFS architecture and MapReduce framework.
* Developed Python jobs to automate transfer of data from Cassandra.
* Implemented the Lambda architecture at enterprise level using technologies like Flume, Kafka. Spark Streaming and Cassandra. The architecture allowed ingestion of data both in the real time and batch setup so that the new incoming records can be used for analytical system built on the batch holders of data.
* Experience on installation and configuration of Spark for Testing and development environments.
* Good knowledge on installation and administration of Cassandra, MongoDB and HBase
* Integrating user data from Cassandra to data in HDFS. Integrating Cassandra with Spark for real time user attributes look up.
* Experience in importing data from a relational database management system (RDBMS) such as MySQL and Oracle into the Hadoop Distributed File System (HDFS), and exporting the processed
* Very Good experience on Unix, Linux scripting
* Support development, testing, and operations teams during new system deployments.
* Evaluate and propose new tools and technologies to meet the needs of the organization.
* Always worked closely with system users and management personnel and gained reputation of loyalty.
* Major strengths are familiarity with multiple software systems, ability to learn quickly new technologies, adapt to new environments, self-motivated, focused, adaptive and quick learner with excellent interpersonal, technical and communication skills.
* About to give AWS Solutions Architect certification

**TechnicalSkills**:

|  |  |
| --- | --- |
| Big Data Ecosystems | Hadoop, MapReduce, HDFS, HBase, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, YARN, Splunk(Log Analytics),Spark, Kafka, Impala,Cloudera CDH 4, CDH5, MapR-M3, MapR-M5, HDP 2.5 ( HORTONWORKS), |
| **Automation** | Puppet, Shell Scripting. |
| Operating Systems | Windows, Linux, UNIX |
| Languages | Scala, Python ,Core Java, R |
| cloud | AWS |
| Databases | MySQL 4.x/5.x, Oracle-9i, Oracle-10G,Oracle 11G, DB2, Netezza |
| NoSQL Technologies | Cassandra,MongoDB, HBase |
| Reporting tools | Tableau, Qlikview |
| ETL | Datastage, Pentaho, Hive |

ACADEMIC QUALIFICATION

Bachelor of Technology (B.Tech) in Electronics and Communications Engineering, JNTU Anantapur, with 87.33 % in 2010

ACHIEVEMENTS

**JNTU Anantapur GOLD Medalist In** Electronics and Communications Engineering, JNTU **ECE (2006-2010)**

BIGDATA CERTIFICATIONS

* MapR Certified Hadoop Developer (MCHD )(5.1: Certificate no: s8cfskwqbruf)
* MapR Certified Cluster Administrator 5.1(Certificate no: n4994n35d9sf)
* Cloudera Certified Developer for Apache Hadoop (CCDH) (License: 100-010-314)
* AWS Solutions Architect- Associate

PROFESSIONAL WORK EXPERIENCE

**PROJECTS COMENCED**

**Pekin Insurance /SWIFT TECH LLC 16th Oct 2017- Now**

**Lead Hadoop Engineer/ Solutions Architect**

**Description : Data Management and Hortonworks Hadoop DataLake Design**

* Analyze, design, architect and coordinate the development of subject area data mart systems to support enterprise information needs, and reporting solutions.
* Designing and deploying scalable, highly available, and fault tolerant systems on AWS
* Lift and shift of an existing on-premises application to AWS
* Ingress and egress of data to and from AWS
* Selecting the appropriate AWS service based on data, compute, database, or security requirements
* Identifying appropriate use of AWS architectural best practices
* Estimating AWS costs and identifying cost control mechanisms
* Spearheading technology discussions to evaluate new tools, technologies, and strategic partnerships.
* Experience in leading advanced analytics solutions, identifies opportunities to acquire new data sources
* Performance tuning of various components in HDP to get the effective utilization of cluster resources
* Designed Spark schema and data selection queries and Involved in data ingestion process.
* Created data frames and tables in Spark for low latency and In-memory querying.

**Loan Genius/SWIFT TECH LLC**

**Lead Hadoop Consultant Aug 2017 - 13th Oct 2017**

**Description : DataLake Design consultation**

**Design and Architectural consultation for Finance platform(Design & Solution Implementation)**

**CGI**

**Client: Scotia Bank, Canada Feb 2017 to Jul 2017**

**Lead Hadoop Analyst/ Solutions Architect**

**Description: Building Enterprise level Data Lake and Processing Data(EDL+Streaming+Analytics)**

The goal of this project is to build data lake for capturing different subject area files that is arriving into Scotia Network through different vendors. Main objective here is to capture the data , processing and doing transformations on go with the data captured and processing into Final Consumer zone. The Final data will be available for Consumer analytics zone for analytics and reporting. The Primary area of the project mainly focused on Building and set up of Data Lake and then loading and processing of data with high efficient data driven platforms like Hadoop and its associated streaming architectures.

The Data Lake is a data-centered architecture featuring a repository capable of storing vast quantities of data in various formats. Data from web server logs, data bases, social media, and third-party data is ingested into the Data Lake. Security policies, including entitlements, also are applied. The main benefit of a data lake is the **centralization** of disparate content sources. Once gathered together (from their “information silos”), these sources can be combined and processed using big data, search and analytics techniques which would have otherwise been impossible. The disparate content sources will often contain proprietary and sensitive information which will require implementation of the appropriate security measures in the data lake

**Accountabilities:**

* Performance tuning of various components in HDP to get the effective utilization of cluster resources
* Designed Spark schema and data selection queries and Involved in data ingestion process.
* Created data frames and tables in Spark for low latency and In-memory querying.
* Involved in spark jobs performance tuning.
* Observing the logs and performing log analytics framework using SPLUNK
* Extensive experience in developing jars with SBT build tool and also Scala ide using SBT library or maven dependencies for integrating Spark with Kafka, Spark with Flume, Spark with HDFS and Hive then store the queried data to HDFS, HBase, and Cassandra.
* Developed custom Kafka producer and consumer for different publishing and subscribing to Kafka brokers.
* Pushed data to and fro from Cassandra into spark for quick Data refresh.
* Connected Spark and Hive tables for interoperability
* Experience in developing application, automated scripts leveraging MongoDB.
* Setting up with Capacity Scheduler and Core site Configuration files for the Multi-tenancy check.
* Stored and Processed structured, semi-structured and unstructured data in Data-Lake.
* Developing and Running Example programs in Scala and Java with Spark and Hadoop Frame-work in Data-Lake.
* Code deployment in Amazon EMR using AWS management tools(CLI).
* Implemented the Lambda architecture using Kafka, Spark and Cassandra so as to maintain an ecosystem of data at rest and data in motion.
* Writing Scripts for loading data into **MongoDB** and involved in backup and restore activities
* Involved in review of functional and non-functional requirements.
* Installed and configured Hadoop **Mapreduce**, **HDFS**, Developed multiple MapReduce jobs in java for data cleaning and preprocessing.
* Importing and exporting data into **HDFS** and **Hive** using **Sqoop**.
* Worked in running **Hadoop** streaming jobs to process terabytes of xml format data.
* Load and transform large sets of structured, semi **structured** and **unstructured** data.
* Designed and Implemented Real time applications using **Apache Storm**, **Kafka**, and **Accumulo**.
* Tested **Map Reduce** programs using MR unit.
* Involved in loading data from **UNIX** file system to **HDFS**.
* Installed and configured **Hive** and also written **HiveUDFs**.
* Configured authorization to Cassandra cluster using Password Authenticator and Kerberos for Hadoop cluster.
* Performed importing data from various sources to the Cassandra cluster using Java/Python APIs.

**Environment:HDP2.5(Hortonworks)**,Kafka, Spark, Cassandra, Hive, Tableau, Scala for Spark, Mysql, Flume.

**Datamatics Global Services Ltd**

**Client - Bank of America , New York , NY Dec 2015 – Dec 2016**

**Description: Credit Payment fraud detection**

**Role : Hadoop Big Data - Lead Consultant**

As per VISA INC there is a fraud of million dollars annually. VISA, being the one of the top Multinational Financial services in the world it processes more than millions of financial transactions daily and there is every possibility that some of them are fraudulent transactions. The goal of this project is to identify possible fraud transactions out of the total financial transaction processed daily.

**Accountabilities:**

* Gained very good business knowledge on financial transactions processing, fraud suspect identification, appeals process etc.
* Installed and configured **Hadoop Map Reduce**, **HDFS**, Developed multiple maps reduce jobs in java for data cleaning and preprocessing.
* Getting online streaming data using **Spark APIs**
* Involved in loading data from UNIX file system to HDFS.
* Installed and configured Hive and also written Hive UDFs.
* Writing Scripts for loading data into **MongoDB** and involved in backup and restore activities
* Prepare an environment for Hadoop cluster setup by installing all the prerequisites.
* Responsible for maintaining data and development in APACHE HADOOP.
* Involved in analyzing and debugging errors occurring during jobs execution in Big Data cluster environment
* Developed Hive Databases, Tables, Partition, Maps and Arrays.

**Datamatics Global Services Ltd**

**Client - BOA , New York, NY April 2014 – Dec 2015**

**Description: Document Storage & Retrieval System**

**Hadoop Big Data - Lead Consultant**

Document Storage & Retrieval System classifies the incoming document using Naive Bayes classifier for efficient storage. The Classifier will be trained with training documents and based on the training information the classifier will classify the incoming documents of different format (doc, docx, pdf, txt, csv) and store them in document oriented database – Mongo DB. The metadata information also will be extracted and stored in Mongo DB. The metadata information helps in the retrieval process of the documents stored in Mongo DB. The documents information will be queried through GUI and based on the results necessary documents will be retrieved.

Here the Purpose of Big data framework Hadoop Technology used to process the storage of classification documents Parallel. So that user would get the efficient required document in a right time by just providing few inputs like any Information of document.

We have built 20 Node MapR cluster and deployed application and checked the performance metrics.

**Accountabilities:**

* Hands on experience in installation, configuration, supporting and managing Hadoop Clusters using Apache, MapR and MongoDB Clusters
* Coding and testing the classifier functionality using java programming
* Bring near real-time data into HDFS.
* Involved in migration of projectfrom Hadoop to Spark
* Use Hive to explicitly control the map- reducetransform stages.
* WritingHive UDF and itsintegrationwith the application
* Involved in writing Hive scripts to extract, transform and load the data into Database.
* Worked on setting up Hive on multiple nodes and developed using Hive.
* Involved in developing the HIVE queries, developing Reports and charts
* Created multiple databases with sharded collections and choosing shard key based on the requirements.
* Testing, evaluation and troubleshooting of different NoSQL database systems and cluster configurations to ensure high-availability in various crash scenarios.
* Upgraded Mongo systems and involved with creating shards, replica sets, monitoring, and projections for Mongo Systems.
* Developed Ruby/Python scripts to monitor health of Mongo databases and perform ad-hoc backups using **Mongodump** and **Mongorestore**.
* Used MongoDB third party tools (MongoChef, MONGBOOSTER, MONGOVUE) and mongo-built in binaries to monitor and analyze the performance of MONGODB.

Worked with capped collections to wipe out aged data based on requirements

**Datamatics Global Services Ltd**

**Client - PCH ,Long Islands, New Jersey Feb 2013 – April 2014**

**Description:XML Processing with Map Reduce and Hive**

**Hadoop Big Data - Senior Software Engineer : Big Data Labs**

The BARC team to store and analyse data from TV meters.ProposedBig Data and Analytics based**BARC-TAM VALIDATION PLATFORM** to provide the followingbenefits.

* Bring near real-time data into HDFS.
* Effectivelydecouples computation fromstoragethankey-value servers
* Easierscale out architecture.
* Tolerating hot spots (resultingfrom MR jobs) easily
* Use Hive to explicitly control the map- reducetransform stages.
* Hivequeriesbuiltcanbeconnected to several Hadoop packages for statisticalprocessing.

**Accountabilities:**

As a team member, was responsible for

* Project requirement analysis and designing architectural flow of the process.
* Setting up the environment including configuration of Mongo DB in multi- nodes cluster and Designing GUI for user interaction.
* Have implemented Distributed Cache in Map-Reduce.
* Implemented Map-Reduce program to extract the data from xml to structured format.
* Loading data from HDFS to hive.
* Have worked with internal tables and External tables for data manipulations.

**Textron IT Services**

**Client - FORD Bangalore, India March 2012 – Feb 2013**

**Description: Automatic Parking System (APS), Online Forum (OF)**

**Hadoop Big Data - Senior Software Engineer**

**Accountabilities:**

* Project requirement analysis and designing architectural flow of the process.
* Have implemented Distributed Cache in Map-Reduce.
* Implemented Map-Reduce program to extract the data from xml to structured format.
* Loading data from HDFS to hive
* Have worked with internal tables and External tables for data manipulations

**Textron IT Services**

**Client: AMEX Bangalore, India Jan 2010 – March 2012**

**Description** :**PRMS/CARS, BARX, SCM, COMMODITIES, GOS**

**Assistant Software Engineer**

AMEX is a major investment bank involved in equity, bond and fund management. AMEX has around 200 applications with different technologies running on Sybase database. It requires all the applications to be converted to run on a new database – either SQL Server/Oracle. As the applications are spread over multiple technologies, expertise on multiple technologies and databases is required to complete projects on time. Textron can help AMEXs capital complete these projects by providing expertise in all the required technologies. Textron will study the existing systems, frame the migration strategy, migrate the databases, re-write the existing applications, test them and deliver them to SBI's capital

**Accountabilities:**

* Migrate the database objects to Oracle from Sybase
* Handling the Data loading and Data validation for the entire databases
* Was able to handle the deliveries to client to be accurate and efficient in time.
* Developed automation Scripts for Data loading and Data validation to optimize the time.
* Coordinate with client on the progress of the project
* Was a part of System Testing, System Integration Testing phases and worked on Autosys jobs in all the Projects in these phases.
* Provided Complete support while GO LIVE of the project and handled issues effectively