**Eric Thurston**

**Eric.thurston@outlook.com**

**Professional Summary:**

* 8+ years of professional IT work experience in analysis, design, development, testing and implementation of Hadoop, Bigdata Technologies like Hadoop and spark ecosystems, Data Warehousing, and AWS on Object Oriented Programming.
* Having 4+ years of comprehensiveexperience in Bigdata using Hadoop and its ecosystem components like HDFS, Spark with Scala and python, Zookeeper, Yarn, MapReduce, Pig, Sqoop, HBase, Hive, Flume, Oozie, Kafka, Flume, Spark streaming and TEZ.
* Worked on NoSQL databases like MongoDB,HBase,Cassandra.
* Experience in Data Modeling and working with Cassandra Query Language (CQL).
* Experience on using Cassandra CQL with Java APIs to retrieve data from Cassandra tables.
* Hands on experience in querying and analyzing data from Cassandra for quick searching, sorting and grouping through CQL.
* Experience in implementing spark solution to enable real time reports from Cassandra data.
* Hands on expertise in working and designing of Row keys & Schema Design with NOSQL databases like Mongo DB.
* Experience in extracting files from MongoDB through Sqoop and placed in HDFS and processed.
* Experience with NoSQL database by using Indexing, Replication and Sharding in MongoDB. Sorted the data by using indexing.
* Experienced with performing CRUD operations using HBase Java Client API.
* Expertise in implementing Ad-hoc queries using Hive QL and good knowledge in creating Hive tables and loading and analyzing data using hive queries.
* Experienced in working with structured data using HiveQL, join operations, Hive UDFs, partitions, bucketing and internal/external tables.
* Experienced in using Pig scripts to do transformations, event joins, filters and some pre-aggregations before storing the data onto HDFS.
* Involved in Debugging Pig and Hive scripts and used various optimization techniques in MapReduce jobs. Wrote custom UDFs and UDAF for Hive and Pig core functionality.
* Worked on relative ease with different working strategies like Agile, Waterfall, Scrum, and Test-Driven Development(TDD) methodologies.
* Experience with all stages of the SDLC and Agile Development model right from the requirement gathering to Deployment and production support.
* Hands on experience with AWS components like EC2, S3, Data Pipeline, RDS, RedShift and EMR.
* Imported the data from different sources like AWSS3, Local file system into Spark RDD and worked on cloud Amazon Web Services (EMR, S3, EC2, Lambda).
* Hands on experience in working with Flume to load the log data from multiple web sources directly into HDFS.
* Experience in importing and exporting data using Sqoop from RelationalDatabase Systems to HDFS.
* Uploaded and processed terabytes of data from various structured and unstructured sources into HDFS (AWS cloud) using Sqoop.
* Had a very good exposure working with various File-Formats (Parquet, Avro & JSON) and Compressions(Snappy &Gzip).
* Hands on experience in creating and designing data ingest pipelines using technologies such as Apache Storm- Kafka.
* Good working experience on Spark (spark streaming, spark SQL) with Scala and Kafka. Worked on reading multiple data formats on HDFS using Scala.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Hands on experience with Spark Core, Spark SQL and Data Frames/Data Sets/RDD API.
* Replaced existing map-reduce jobs and Hive scripts with Spark Data-Frame transformation and actions. Good knowledge on Spark architecture and real-time streaming using Spark with Kafka.
* Experienced working with Spark Streaming, SparkSQLand Kafka for real-time data processing.
* Created dataflow between SQL Server and Hadoop clusters using Apache Nifi.
* Experienced in running query using Impala and used BI tools to run ad-hoc queries directly on Hadoop.
* Experience in using Kafka and Kafka brokers to initiate spark context and processing livestreaming
* informationwith the help of RDD.
* Experience in developing and scheduling ETL workflows in Hadoop using Oozie with the help of deployment and managing Hadoop cluster using Cloudera and Horton works.
* Used Oozie and Zookeeper operational services for coordinating cluster and scheduling workflows.
* Experience with version control tools like Git, CSV and SVN.
* Excellent experience in designing and developing Enterprise Applications for J2EE platform using Servlets, JSP, Struts, Spring, Hibernate and Web services.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big data/Hadoop** | **Hadoop 2.7/2.5, HDFS 1.2.4, Map Reduce, Hive, Pig, Sqoop, Oozie, Hue.** |
| **NoSQL Databases** | **HBase, MongoDB3.2 & Cassandra** |
| **Programming Languages** | **Java, Python, SQL, PL/SQL, AWS, Hive QL, Unix Shell Scripting, Scala** |
| **IDE and Tools** | **Eclipse 4.6, Netbeans 8.2, BlueJ** |
| **Web Technologies** | **HTML 5/4, DHTML, AJAX, JavaScript, jQuery and CSS3/2, JSP, Bootstrap 3/3.5** |
| **Application Server** | **Apache Tomcat, Jboss, IBM Web sphere, Web Logic** |
| **Java/J2EE Technologies** | **Servlets, JSP, JDBC, JSTL, EJB, JAXB, JAXP, JMS, JAX-RPC, JAX- WS** |
| **Operating Systems** | **Windows 8/7, UNIX/Linux and Mac OS.** |
| **Database** | **Oracle 12c/11g, MYSQL, SQL Server 2016/2014** |
| **Other Tools** | **Maven, ANT, WSDL, SOAP, REST.** |
| **Methodologies** | **Waterfall, Agile UML, Design Patterns (Core Java and J2EE)** |

**CISCO, San Jose, CA Jun 2017 – Till Date**

**Big Data Developer**

**Responsibilities:**

* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Working on Big Data infrastructure for batch processing as well as real-time processing. Responsible for building scalable distributed data solutions using Hadoop.
* Implemented apache Airflow DAG to find popular items in redshift and ingest in the main PostgreSQL via a web service call.
* Worked on Spark and MLlib to develop a linear regression model for logistic information.
* Involved in designing Kafka for multi data center cluster and monitoring it.
* Responsible for importing real time data to pull the data from sources to Kafka clusters.
* Responsible for design and development of Spark SQL Scripts based on Functional Specifications.
* Working experience on Spark ecosystems using spark components like Spark Core, Spark SQL, Spark Streaming, MLlib,and GraphX.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, and Scala.
* Developed Spark Applications by using Scala, Java, Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Experience with developing and maintaining Applications written for Amazon Simple Storage, AWS Elastic Beanstalk, and AWS Cloud Formation.
* Used Spark SQL on data frames to access hive tables into spark for faster processing of data.
* Working knowledge of Spark RDD, Data Frame API, Data set API, Data Source API, Spark SQL and Spark Streaming.
* Used Different Spark Modules like Spark core, Spark SQL, Spark Streaming, Spark Data sets and Data frames.
* Responsible for developing multiple Kafka Producers and Consumers from scratch as per the software requirement specifications.
* Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Worked on the large-scale Hadoop YARN cluster for distributed data processing and analysis using Spark, Hive.
* Involved in creating data-lake by extracting customer's data from various data sources to HDFS which include data from Excel, databases, and log data from servers.
* Created various hivemanaged tables, staging tables and joined the tables as per the requirement. Implemented static Partitioning, Dynamic partitioning and Bucketing in Hive using internal and external table.
* Moved Relational Database data using Sqoop into Hive Dynamic partition tables using staging tables.
* Involved in collecting, aggregating and moving data from servers to HDFS using Apache Flume.
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
* Successfully migrated the data from AWS S3 source to the HDFS sink using Kafka.
* Queried and analyzed data from Cassandra for quick searching, sorting and grouping through CQL.
* Involved in performing importing data from various sources to the Cassandra cluster using Sqoop.
* Worked on creating data models for Cassandra from Existing Oracle data model.
* Designed Column families in Cassandra and Ingested data from RDBMS, performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Reading the log files using Elastic Search Logstash and alerting users on the issue and saving the alert details to Cassandra for analyzations.
* Used Impala where ever possible to achieve faster results compared to Hive during data Analysis.
* Experienced in running query using Impala and used BI tools to run ad-hoc queries directly on Hadoop.
* Worked extensively on Apache NiFi to build Nifi flows for the existing oozie jobs to get the incremental load, full load, semi structured data and to get data from rest API into Hadoop and automate all the Nifi flows runs incrementally.
* Created Nifi flows to trigger spark jobs and used put email processors to get notifications if there are any failures.
* Implemented Apache Nifi flow topologies to perform cleansing operations before moving data into HDFS.
* Used git as version control tool to update work process.
* Implemented and maintained the monitoring and alerting of production and corporate servers/storage using AWS Cloud watch.
* Worked on apacheSolr for indexing and load balanced querying to search for specific data in larger datasets
* Implemented the workflows using Apache Oozie framework to automate tasks. Used Zookeeper to co-ordinate cluster services.
* Experience in using version control tools like GITHUB to share the code snippet among the team members.
* Involved in daily SCRUM meetings to discuss the development/progress and was active in makingscrum meetings more productive.

**Environment:** Hadoop, Map Reduce, HDFS, Hive, Cassandra, Sqoop, Oozie, SQL, Kafka, Spark, Scala, Java, AWS, GitHub, Solr, Impala.

**Inovalon, Herndon, VA Sep 2015– May 2017**

**Big Data/Hadoop Developer**

**Responsibilities:**

* Built scalable distributed Hadoop cluster running Hortonworks Data Platform.
* Develop data set processes for data modelling, and mining. Recommend ways to improve data reliability, efficiency and quality.
* Working experience with data streaming process with Kafka, Apache Spark, Hive, Pig, etc.
* Importing and exporting data into HDFS Sqoop and Flume and Kafka.
* Utilized Flume to filter out the input data read to retrieve only the data needed to perform analytics by implementing flume interception.
* Created data pipelines for different events to load the data from DynamoDB to AWS S3 bucket and then into HDFS location.
* Involved in collecting, aggregating and moving data from servers to HDFS using Apache Flume.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the data got from Kafka.
* Developed a Nifi Workflow to pick up the data from SFTP server and send that to Kafka broker
* Worked on analysing Hadoop Cluster and different big data analytic tools including Pig, Hive.
* Extracted files from MongoDB through Sqoop and placed in HDFS for processed.
* Configured Flume to extract the data from the web server output files to load into HDFS.
* Used Flume to collect, aggregate and store the web log data from different sources like web servers, mobile and network devices and pushed into HDFS.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Analysed the SQL scripts and designed the solution to implement using Scala.
* Used Spark-SQL to Load JSON data and create SchemaRDD and loaded it into Hive Tables and handled structured data using Spark SQL.
* Tested Apache Tez for building high performance batch and interactive data processing applications on Pig and Hive jobs.
* Implemented messaging queue to integrate with Cassandra using Apache Kafka and Zookeeper.
* Implemented Real time analytics on Cassandra data using thrift API.
* Designed Columnar families in Cassandra and Ingested data from RDBMS, performed transformations and exported the data to Cassandra.
* Queried and analyzed data from Cassandra for quick searching, sorting and grouping through CQL.
* Developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with Pig.
* Involved in complete Big Data flow of the application starting from data ingestion upstream to HDFS, processing the data in HDFS and analysing the data and involved.
* Worked on Apache Nifi to decompress and move JSON files from local to HDFS.
* Developed and designed automate process using shell scripting for data movement.
* Involved in loading data from UNIX file system to HDFS using ShellScripting.

**Environment:** Java, J2EE 1.7, Eclipse, Apache Hive, HDFS, Github, Jenkins, NiFi, Python, Scala, Pig, Hadoop, Scripting and AWS S3, EC2, Impala, Shell Scripting, Apache Web Server, Spark, Spark SQL, JIRA.

**Texas Instruments - Dallas, TX Oct 2013- Aug 2015**

**Hadoop Developer**

**Responsibilities:**

* Responsible for building scalable distributed data solutions using Hadoop.
* Hadoop installation, Configuration of multiple nodes using Clouder platform.
* Installed and configured a Hortonworks HDP 2.2 using Ambari and manually through command line. Cluster maintenance as well as creation and removal of nodes using tools like Ambari, Cloudera Manager Enterprise and other tools.
* Handling the installation and configuration of a Hadoop cluster.
* Building and maintaining scalable data pipelines using the Hadoop ecosystem and other open source components like Hive and HBase.
* Involved in developer activities of installation and configuring Hadoop MapReduce, HDFS, Developed multiple MapReduce jobs in java for data cleaning and preprocessing.
* Importing and exporting data into HDFS and Hive using Sqoop.
* Involved in Cluster Level Security, Security of perimeter (Authentication- Cloudera Manager, Active directory and Kerberos) Access (Authorization and permissions- Sentry) Visibility (Audit and Lineage - Navigator) Data ( Data Encryption at Rest) ➢ Handling the data exchange between HDFS and different web sources using Flume and Sqoop.
* Monitoring the data streaming between web sources and HDFS and functioning through monitoring tools.
* Close monitoring and analysis of the MapReduce job executions on cluster at task level.
* Inputs to development regarding the efficient utilization of resources like memory and CPU utilization based on the running statistics of Map and Reduce tasks.
* Install OS and administrated Hadoop stack with CDH5 (with YARN) Cloudera Distribution including configuration management, monitoring, debugging, and performance tuning Scripting Hadoop package installation and configuration to support fully-automated deployments.
* Day-to-day operational support of our Cloudera Hadoop clusters in lab and production, at multi-petabyte scale.
* Changes to the configuration properties of the cluster based on volume of the data being processed and performed by the cluster.
* Involved in creating Spark cluster in HDInsight by create Azure compute resources with Spark installed and configured.
* Setting up automated processes to analyze the system and Hadoop log files for predefined errors and send alerts to appropriate groups and an Excellent working knowledge on SQL with databases.
* Commissioning and De-commissioning of data nodes from cluster in case of problems.
* Setting up automated processes to archive/clean the unwanted data on the cluster, in particular on Name Node and Secondary Name node.
* Set up and managing HA Name Node to avoid single point of failures in large clusters.
* Discussions with other technical teams on regular basis regarding upgrades, process changes, any special processing and feedback.
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions. Documented the systems processes and procedures for future references.
* Worked with systems engineering team to plan and deploy new Hadoop environments and expand existing Hadoop clusters.
* Administering and Maintaining Cloudera Hadoop Clusters Provision physical Linux systems, patch, and maintain them.

**Environment**: Hadoop, Confluent Kafka, Hortonworks HDF, HDP, NIFI, Linux, Splunk, Yarn, Clouder 5.13, Spark, Tableau.

**Fannie Mae - Herndon, VA June 2011-Sep 2013**

**Hadoop Admin**

**Responsibilities:**

* Launching Amazon EC2 Cloud Instances using Amazon Web Services (Linux/ Ubuntu/RHEL) and Configuring launched instances with respect to specific applications.
* Installed application on AWS EC2 instances and also configured the storage on S3 buckets.
* Performed S3 buckets creation, policies and also on the IAM role based polices and customizing the JSON template.
* Managed servers on the Amazon Web Services (AWS) platform instances using Puppet, Chef Configuration management.
* Developed PIG scripts to transform the raw data into intelligent data as specified by business users.
* Worked in AWS environment for development and deployment of Custom Hadoop Applications.
* Worked closely with the data modelers to model the new incoming data sets.
* Involved in start to end process of Hadoop jobs that used various technologies such as Sqoop, PIG, Hive, Map Reduce, Spark and Shell scripts (for scheduling of few jobs.
* Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, Oozie, Zookeeper, SQOOP, flume, Spark, Impala, Cassandra with Horton work Distribution.
* Involved in creating Hive tables, Pig tables, and loading data and writing hive queries and pig scripts. Assisted in upgrading, configuration and maintenance of various Hadoop infrastructures like Pig, Hive, and HBase.
* Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data. Configured deployed and maintained multi-node Dev and Test Kafka Clusters.
* Performed transformations, cleaning and filtering on imported data using Hive, Map Reduce, and loaded final data into HDFS.
* Worked on tuning Hive and Pig to improve performance and solve performance related issues in Hive and Pig scripts with good understanding of Joins, Group and aggregation and how it does Map Reduce jobs testing and processing of data.
* Import the data from different sources like HDFS/HBase into Spark RDD.
* Developed a data pipeline using Kafka and Storm to store data into HDFS.
* Performed real time analysis on the incoming data.
* Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing.
* Implemented Spark using Scala and SparkSQL for faster testing and processing of data.

**Environment**: Apache Hadoop, HDFS, MapReduce, Sqoop, Flume, Pig, Hive, HBASE, Oozie, Scala, Spark, Linux.