

Cesar Niyomugabo

Data Engineer

Email: consultant@gmail.com

Phone: 999-999-9999

Professional Profile

- 7 years of experience in development of custom Hadoop Big Data solutions, platforms, pipelines, data migration, and data visualizations.
- Ability to troubleshoot and tune relevant programming languages like SQL, Java, Python, Scala, PIG, Hive, RDDs, DataFrame & MapReduce. Able to design elegant solutions through the use of problem statements.
- Created classes that simulate real-life objects and write loops to perform actions on your data.
- AWS tools (Redshift, Kinesis, S3, EC2, EMR, DynamoDB, Elasticsearch, Athena, Firehose, Lambda)
- Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems/ Non- Relational Database Systems and vice-versa Accustomed to working with large complex data sets, real-time/near real-time analytics, and distributed big data platforms.
- Experience with multiple terabytes of data stored in AWS using Elastic Map Reduce (EMR) and Redshift
- Developed data queries using HiveQL and optimized the Hive queries
- Expertise in developing PIG Latin Scripts and Hive Query Language for data analytics. Well-versed in and implemented Partitioning, Dynamic-Partitioning and bucketing concepts in Hive to compute data metrics.
- Strong knowledge of Pig and Hive's analytical functions, extending Hive and Pig core functionality by writing custom UDFs.
- Experience in developing REST API's for use in single page or native applications and
- Created Hive Managed and External tables with partition and bucket in Hive and loaded data in to Hives
- In-depth understanding/knowledge of Hadoop Architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and MapReduce concepts and experience in working with MapReduce programs using Apache Hadoop for working with Big Data to analyze large datasets efficiently.
- Excellent Knowledge in understanding Big Data infrastructure, distributed file systems -HDFS, parallel processing - MapReduce framework and complete Hadoop ecosystem - Hive, Hue, Pig, HBase, Zookeeper, Sqoop, Kafka-Storm, Spark, Flume, and Oozie.
- In-depth knowledge of real-time ETL/Spark analytics using Spark SQL with visualization Hands-on experience on YARN (MapReduce 2.0) architecture and components such as Resource Manager, Node Manager, Container and Application Master and execution of a MapReduce job.

Cesar Niyomugabo

Data Engineer

Email: consultant@gmail.com

Phone: 999-999-9999

Professional Skills

EXPERIENCE

10 years of experience in the field of data analytics, data processing and database technologies.

7 years of experience with the Hadoop ecosystem and Big Data tools and frameworks.

PROJECT MANAGEMENT

Agile, Kanban, Scrum, DevOps, Continuous Integration, Test-Driven Development, Unit Testing, Functional Testing, Design Thinking, Lean, Six Sigma

DATABASE

SQL, NoSQL, Apache Cassandra, MongoDB, Hbase, RDBMS, Hive

SOFTWARE

AutoCAD • MATLAB • Revit • LTspice • PSpice • Multisim • Microsoft Office Suites

BIG DATA PLATFORMS

Amazon AWS, Microsoft Azure, Elasticsearch, Apache Solr, Lucene, Cloudera Hadoop, Cloudera Impala, Databricks, Hortonworks Hadoop

PROGRAMMING

Python, Scala, PHP • Python • Bash • LISP • SQL • JavaScript • JQuery • C • C++ • XML • HTML • CSS, Visual Basic, VBA, .Net, Spark, HiveQL, Spark API, REST API

DATA VISUALIZATION

Tableau, Microsoft Power BI

FILES

HDFS, Avro, Parquet, Snappy, Gzip, SQL, Ajax, JSON, GSON, ORC

OPERATING SYSTEMS

Linux, MacOS, Microsoft Windows

HADOOP ECOSYSTEM COMPONENTS & TOOLS

Apache Ant, Apache Cassandra, Apache Flume, Apache Hadoop, Apache Hadoop YARN, Apache Hbase, Apache Hcatalog, Apache Hive, Apache Kafka, Apache MAVEN, Apache Oozie, Apache Pig, Apache Spark, Spark Streaming, Spark MLlib, GraphX, SciPy, Pandas, RDDs, DataFrames, Datasets, Mesos, Apache Tez, Apache ZooKeeper, Airflow and Camel, Apache Lucene, Elasticsearch, Apache SOLR, Apache Drill, Presto, Apache Hue, Sqoop, Kibana,

Cesar Niyomugabo

Data Engineer

Email: consultant@gmail.com

Phone: 999-999-9999

Professional Experience

DATA ENGINEER

January 2019 – Present

American Red Cross

Washington, D.C

-
- Used Hadoop cluster to manage and perform data ingestion from Rapid API
 - Created and maintained a cluster of multiple Kafka brokers to ingest data from Kafka producer
 - Used spark to build and process real-time data stream from Kafka producer
 - Used Spark DataFrame API over Cloudera platform to perform analytics on data.
 - Defined and implemented schema for a custom Hbase
 - Used SparkSQL for creating and populating hbase warehouse
 - Executed Hadoop/Spark jobs on AWS EMR using programs, data stored in S3 Buckets.
 - Added support for Amazon AWS S3 and RDS to host static/media files and the database into Amazon Cloud.
 - Implemented advanced procedures of feature engineering for data science team using the in-memory computing capabilities like Apache Spark written in Scala
 - Wrote streaming applications with Spark Streaming/Kafka.
 - Used SparkSQL module to store data into HDFS
 - Configured Kafka broker for the Kafka cluster of the project and streamed the data to Spark for structured streaming to get structured data by schema.
 - Used Spark DataFrame API over Cloudera platform to perform analytics on data.
 - Handled over millions of messages per a day funneled through Kafka topics.
 - Worked with Jenkins CI for CICD and Git version control.
 - Optimized ETL jobs to reduce memory and storage consumption.
 - Communicated and present findings, orally and visually in a way that can be easily understood by business counterparts

Cesar Niyomugabo

Data Engineer

Email: consultant@gmail.com

Phone: 999-999-9999

DATA ENGINEER

October 2017 - January 2019

SIEMENS

Princeton, NJ

Project Description

- Engaged constructively with project teams to support project objectives through the application of sound architectural principles
- Configured flume agent source, sink and channel for data stream collection on API
- Used flume for collection and ingestion of data from API to HDFS
- Integrated flume with Spark Streaming for real time data processing
- Used spark to load and process data from HDFS
- Used Sqoop to export data from HDFS to MYSQL database for deep analysis queries
- Developed POC using Scala & deployed on Yarn cluster, compared the performance of Spark, with Hive and SQL
- Used hive for queries and incremental imports with Spark and Spark jobs for data processing and analytics
- Installed and configured Kafka cluster and monitoring the cluster; Architected a lightweight Kafka broker; integration of Kafka with Spark for real time data processing
- Built a Spark proof of concept with Python using PySpark
- Implemented advanced procedures of feature engineering for data science team using the in-memory computing capabilities like Apache Spark written in Scala
- Extracted the needed data from the server into Hadoop file system (HDFS) and bulk loaded the cleaned data into HBase using Spark
- Demonstrated ability to think strategically about business, product, and technical challenges in an enterprise environment
- Used park SQL to create real-time processing of structured data with Spark Streaming processed through structured streaming

Cesar Niyomugabo

Data Engineer

Email: consultant@gmail.com

Phone: 999-999-9999

DATA CLOUD ENGINEER

March 2015 - October 2017

NCR

Atlanta, GA

-
- Created and maintained a cluster of multiple Kafka brokers to ingest data from Kafka producer
 - Used spark to build and process real-time data stream from Kafka producer
 - Used Spark DataFrame API over Cloudera platform to perform analytics on data.
 - Defined and implemented schema for a custom Hbase
 - Used SparkSQL for creating and populating hbase warehouse
 - Executed Hadoop/Spark jobs on AWS EMR using programs, data stored in S3 Buckets.
 - Added support for Amazon AWS S3 and RDS to host static/media files and the database into Amazon Cloud.
 - Wrote streaming applications with Spark Streaming/Kafka.
 - Used sparkSQL module to store data into HDFS
 - Configured Kafka broker for the Kafka cluster of the project and streamed the data to Spark for structured streaming to get structured data by schema.
 - Used Spark DataFrame API over Cloudera platform to perform analytics on data.
 - Worked with Jenkins CI for CICD and Git version control.
 - Optimized ETL jobs to reduce memory and storage consumption.

DATA ENGINEER

September 2013 - March 2015

Ahold

Salisbury, NC

-
- Implemented applications on Hadoop/Spark on Kerberos secured cluster.
 - Responsible for performance optimization of clusters.
 - Set-up Hortonworks Infrastructure from configuring clusters to Node security using Kerberos.
 - Monitored Hadoop cluster using tools like Ambari
 - Migrated the data using Sqoop from HDFS to Relational Database System.

Cesar Niyomugabo

Data Engineer

Email: consultant@gmail.com

Phone: 999-999-9999

- Installed Oozie workflow engine to run multiple Hive Jobs.
- Deep understanding and implementations of various methods to load HIVE tables from HDFS and Local File System.
- Implemented Capacity Schedulers on the Yarn Resource Manager to share the resources of the cluster for the jobs given by the users.

Education

M.S, Machine Learning

Georgia Institute Technology,
Atlanta, USA

M.S, Image Engineering

Chung-Ang University,
Seoul, South Korea

B.S, Electronics and Communication Systems Engineering

National University of Rwanda

CERTIFICATIONS

Microsoft

- Microsoft Professional Program in Artificial Intelligence
- Data Science Research Methods
- Ethics and Law in Data and Analytics
- Essential Mathematics for Machine Learning

Pierian Data INC

- Python for Data Science and Machine learning