MARIO LOPEZ

Hadoop Data Engineer

Profile

Dedicated and seasoned big data professional with skill in implementing and improving big data ecosystems using Hadoop, Spark, Microsoft Azure, Amazon AWS, Cloudera, Hortonworks, MapR, Anaconda, Jupyter Notebooks, and Elastic Search. Proficient in ETL and data pipeline methods and tools.

Professional Summary

- 8 years of experience in the field of data analytics, data processing and database technologies.
- 5 years of experience with the Hadoop ecosystem and Big Data tools and frameworks.
- Ability to troubleshoot and tune relevant programming languages like SQL, Java, Python, Scala, PIG, Hive, RDDs, DataFrames & MapReduce.
- Able to design elegant solutions through the use of problem statements.
- Accustomed to working with large complex data sets, real-time/near real-time analytics, and distributed big data platforms.
- Proficient in major vendor Hadoop distribution like Cloudera, Hortonworks, and MapR.
- Creation of UDF functions in Python or Scala.
- Data Governance, Security & Operations experience.
- Deep knowledge in incremental imports, partitioning and bucketing concepts in Hive and Spark SQL needed for optimization.
- Experience collecting log data from various sources and integrating it into HDFS using Flume; staging data in HDFS for further analysis.
- Experience collecting real-time log data from different sources like webserver logs and social media data from Facebook and Twitter using Flume, and storing in HDFS for further analysis.
- Experience deploying large multiple nodes of a Hadoop and Spark cluster.
- Experience developing custom large-scale enterprise applications using Spark for data processing.
- Experience developing Oozie workflows for scheduling and orchestrating the ETL process.
- Excellent knowledge on Hadoop Architecture and ecosystems such as HDFS, configuration of nodes, YARN, MapReduce, Sentry, Spark, Falcon, Hbase, Hive, Pig, Sentry, Ranger.
- Developed Scripts and automated end-end data management and sync between all the clusters.
- Strong hands on experience in Hadoop Framework and its ecosystem including but not limited to HDFS Architecture, MapReduce Programming, Hive, Pig, Sqoop, HBase, MongoDB, Cassandra, Oozie, Spark RDDs, Spark DataFrames, Spark Datasets, Spark MLlib, etc.
- Involved in building a multi-tenant cluster, with disaster management with Hadoop cluster.
- Experience in Mainframe data and batch migration to Hadoop.
- Hands on experience in installing, configuring Cloudera's and Horton distribution.
- Extending Hive and Pig core functionality by writing custom UDFs.
- Extensively used Apache Flume to collect logs and error messages across the cluster.

MARIO LOPEZ

Hadoop Data Engineer

Technical Skills

Programming Languages & IDEs

■ ■ ■ ■ ■ ■ ■ □ □ □

Unix shell scripting, Object-oriented design, Object-oriented programming, Functional programming, SQL, Java, Hive QL, MapReduce, Python, Scala, XML, Blueprint XML, Ajax, REST API, Spark API, JSON, Avro, Parquet, ORC, Jupyter Notebooks, Eclipse, IntelliJ, PyCharm

DATA & FILE MANAGEMENT

■ ■ ■ ■ ■ ■ ■ ■ ■ □

Apache Cassandra, Apache Hbase, MapR-DB, MongoDB, Oracle, SQL Server, DB2, Sybase, RDBMS, MapReduce, HDFS, Parquet, Avro, JSON, Snappy, Gzip, DAS, NAS, SAN, PostgreSQL, MySQL, MemSQL, OLTP.

Methodologies

■ ■ ■ ■ ■ □ □ □ □ □

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

Cloud Services & Distributions

■ ■ ■ ■ ■ ■ ■ □ □ □

AWS, Azure, Anaconda Cloud, Elasticsearch, Solr, Lucene, Cloudera, Databricks, Hortonworks, Elastic MapReduce

Big Data Platforms, Software, & 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, Cloudera Impala, HDFS, Hortonworks, MapR, MapReduce, Apache Airflow and Camel, Apache Lucene, Elasticsearch, Elastic Cloud, Kibana, X-Pack, Apache

SOLR, Apache Drill, Presto, Apache Hue, Sqoop, Kibana, Tableau, AWS, Cloud Foundry, GitHub, Bit Bucket, Pentaho, Kettle.

Experience

May 2016 **Hadoop Data Architect/Engineer**
Present Comcast Xfinity – Philadelphia, PA

Involved in testing X1 products by gathering data from Comcast X1 Xfinity DVRs, which are used for search and play multimedia content for every subscriber. In exchange, Comcast gets an enormous amount of data - with the idea being that the data can be used to offer the most appropriate content for the user according to their likes and their experience of the platform. For example, selected menu options on the DVRs tell Comcast about utilization, type of content, and geographical variants.

- Involved in creating Hive Tables, loading with data and writing Hive queries, which will invoke and run TEZ jobs in the backend.
- Experience in optimizing the data storage in Hive using partitioning and bucketing mechanisms on both the managed and external tables.
- Worked on importing and exporting data using Sqoop between HDFS to RDBMS.
- Collect, aggregate, and move data from servers to HDFS using Apache Spark & Spark Streaming.
- Administered Hadoop cluster(CDH) and reviewed log files of all daemons.
- Used Impala where possible to achieve faster results compared to Hive during data Analysis.
- Used Spark API over Hadoop YARN to perform analytics on data in Hive.
- Used Spark SQL and DataFrames API to load structured and semi structured data into Spark Clusters.
- Migrated complex MapReduce programs into Apache Spark RDD operations.
- Migrated ETL jobs to Pig scripts for transformations, joins, aggregations before HDFS.
- Implemented data ingestion and cluster handling in real time processing using Kafka.
- Implemented workflows using Apache Oozie framework to automate tasks.
- Performed both major and minor upgrades to the existing Cloudera Hadoop cluster.
- Implemented High Availability of Name Node, Resource manager on the Hadoop Cluster.
- Integrated Hadoop with Active Directory and enabled Kerberos for Authentication.
- Created Hive external tables and designed data models in hive.
- Involved in the process of designing Cassandra Architecture including data modeling.
- Implemented YARN Resource pools to share resources of cluster for YARN jobs submitted by users.
- Performed storage capacity management, performance tuning and benchmarking of clusters.
- Performance tuning of HIVE service for better Query performance on ad-hoc queries.
- Performed performance tuning for Spark Steaming e.g. setting right Batch Interval time, correct level of Parallelism, selection of correct Serialization & memory tuning.
- Data ingestion is done using Flume with source as Kafka Source & sink as HDFS.
- For one of the use case, used Spark Streaming with Kafka & HDFS & Cassandra to build a continuous ETL pipeline. This is used for real time analytics performed on the data.

- Performed import and export of dataset transfer between traditional databases and HDFS using Sqoop.
- Worked on disaster management with Hadoop cluster.
- Designed and presented a POC on introducing Impala in project architecture.
- Configured Spark streaming to receive real time data from Kafka and store the stream data to HDFS.

Environment: HDFS, PIG, Hive, Sqoop, Oozie, HBase, Zoo keeper, Cloudera Manager, java, Ambari, Oracle, MYSQL, Cassandra, Sentry, Falcon, Spark, YARN, MapReduce

May 2015 **Hadoop Data Architect/Engineer**
May 2016 City of Long Beach – Long Beach, CA

The city of Long Beach, California is using smart water meters to detect illegal watering in real time and have been used to help some homeowners cut their water usage by as much as 80 percent. That's vital when the state is going through its worst drought in recorded history and the governor has enacted the first-ever state-wide water restrictions.

- Analyzed Hadoop cluster using big data analytic tools including Kafka, Pig, Hive, Spark, MapReduce.
- Configured Spark streaming to receive real time data from Kafka and store to HDFS using Scale.
- Implemented Spark using Scala and Spark SQL for faster analyzing and processing of data.
- Built continuous Spark streaming ETL pipeline with Spark, Kafka, Scala, HDFS and MongoDB.
- Import/export data into HDFS and Hive using Sqoop and Kafka.
- Involved in creating Hive tables, loading the data and writing hive queries.
- Design and develop ETL workflows using Python and Scala for processing data in HDFS & MongoDB.
- Worked on importing the unstructured data into the HDFS using Spark Streaming & Kafka.
- Wrote complex Hive queries, Spark SQL queries and UDFs.
- Wrote shell scripts to execute scripts (Pig, Hive, and MapReduce) and move the data files to/from HDFS.
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
- Worked with Amazon Web Services (AWS) and involved in ETL, Data Integration and Migration.
- Handled 20 TB of data volume with 120-node cluster in Production environment.
- Loading data from diff servers to AWS S3 bucket and setting appropriate bucket permissions.
- Apache Kafka to transform live streaming with the batch processing to generate reports
- Cassandra data modeling for storing and transformation in spark using Datastax connector.
- Imported data into HDFS and Hive using Sqoop and Kafka. Created Kafka topics and distributed to different consumer applications.
- Worked on Spark SQL and DataFrames for faster execution of Hive queries using Spark and AWS EMR

- Implemented Partitioning, Dynamic Partitions and Buckets in HIVE for increasing performance benefit and helping in organizing data in a logical fashion.
- Scheduled and executed workflows in Oozie to run Hive and Pig jobs
- Worked with Spark Context, Spark -SQL, DataFrame and Pair RDDs.
- Used Hive, spark SQL Connection to generate Tableau BI reports.
- Created Partitions, Buckets based on State to further process using Bucket based Hive joins.
- Created Hive Generic UDF's to process business logic that varies based on policy.
- Developed various data connections from data sourced to SSIS, and Tableau Server for report and dashboard development.
- Worked with clients to better understand their reporting and dash boarding needs and present solutions using structured Waterfall and Agile project methodology approach.
- Developed metrics, attributes, filters, reports, dashboards and also created advanced chart types, visualizations and complex calculations to manipulate the data.

Environment: Hadoop, HDFS, Hive, Spark, YARN, MapReduce, Kafka, Pig, MongoDB, Sqoop, Storm, Cloudera, Impala

Jan 2014 **Hadoop Data Engineer**
May 2015 Gulfstream – Savannah, GA

Offloading Oracle or Teradata Data Warehouses to Hadoop Data Lakes for better scaling, more analytics and cost savings. Created multi-node Hadoop and Spark clusters in AWS instances to generate terabytes of data and stored it in AWS HDFS.

- Deployed the application jar files into AWS instances.
- Used the image files of an instance to create instances containing Hadoop installed and running.
- Developed a task execution framework on EC2 instances using SQL and DynamoDB.
- Designed a cost-effective archival platform for storing big data using Hadoop and its related technologies.
- Connected various data centers and transferred data between them using Sqoop and various ETL tools.
- Extracted the data from RDBMS (Oracle, MySQL) to HDFS using Sqoop.
- Used the Hive JDBC to verify the data stored in the Hadoop cluster.
- Worked with the client to reduce churn rate, read and translate data from social media websites.
- Integrated Kafka with Spark Streaming for real time data processing
- Imported data from disparate sources into Spark RDD for processing.
- Built a prototype for real-time analysis using Spark streaming and Kafka.
- Transferred data using Informatica tool from AWS S3.
- Using AWS Redshift for storing the data on cloud.
- Collected the business requirements from the subject matter experts like data scientists and business partners.
- Involved in Design and Development of technical specifications using Hadoop technologies.
- Load and transform large sets of structured, semi structured and unstructured data.
- Used different file formats like Text files, Sequence Files, Avro.
- Loaded data from various data sources into HDFS using Kafka.
- Tuning and operating Spark and its related technologies like Spark SQL and Streaming.
- Used shell scripts to dump the data from MySQL to HDFS.
- Used NoSQL databases like MongoDB in implementation and integration.
- Worked on streaming the analyzed data to Hive Tables using Sqoop for making it available for visualization and report generation by the BI team.
- Configured Oozie workflow engine scheduler to run multiple Hive, Sqoop and pig jobs.
- Consumed the data from Kafka queue using Storm
- Used Oozie to automate/schedule business workflows which invoke Sqoop, MapReduce and Pig jobs as per the requirements.

Environment: Hadoop, Spark, HDF, Oozie, Sqoop, MongoDB, Hive, Pig, Storm, Kafka, MapReduce, SQL, Acro, RDD. SQS S3, Cloud, MySQL, Informatica, Dynamo DB

Aug 2012 **Hadoop Data Engineer**
Dec 2013 Alibaba – Remote

Involved in building our Data Warehousing solutions for banking industry, pulling data from various sources and file formats.

- Worked with several clients with day to day requests and responsibilities.
- Involved in analyzing system failures, identifying root causes and recommended course of actions.
- Worked on Hive for exposing data for further analysis and for generating transforming files from different analytical formats to text files.
- Wrote the shell scripts to monitor the health check of Apache Tomcat and JBOS; daemon services and respond accordingly to any warning or failure conditions.
- Utilized Java and MySQL from day to day to debug and fix issues with client processes.
- Developed, tested, and implemented financial-services application to bring multiple clients into standard database format.
- Assisted in designing, building, and maintaining database to analyze life cycle of checking and debit transactions.
- Excellent JAVA, J2EE application development skills with strong experience in Object Oriented Analysis, extensively involved throughout Software Development Life Cycle (SDLC).
- Strong experience of software and system development using JSP, Servlet, Java Server Face, EJB, JDBC, JNDI, Struts, Maven, Git, JUnit, SQL language.
- Rich experience of database design and hands-on experience of large database systems: Oracle 8i and Oracle 9i, DB2, PL, SQL.
- Hands-on experience of Sun One Application Server, Web logic Application Server, Web Sphere Application Server, Web Sphere Portal Server, and J2EE application deployment technology.

Environment: Java, JDBC, JNDI, Struts, Maven, Subversion, JUnit, SQL language, spring, Hibernate, JUnit, Oracle, XML, Putty and Eclipse.

Jan 2012 **Hadoop Big Data Administrator**
Dec 2013 Blue Cross Blue Shield of North Carolina – Charlotte, NC

Responsible for administration of big data system using Hadoop on premises including data loss prevention policies, optimization, backup and restore of databases and management of Hadoop clusters on servers.

- Administered Hadoop cluster(CDM) and reviewed log files of all daemons.
- Involved in scheduling Oozie workflow engine to run multiple Hive, Sqoop and Pig jobs.
- Implemented workflows using Apache Oozie framework to automate tasks.
- Used Zookeeper for various types of centralized configurations, GIT for version control, and Maven as a build tool for deploying the code.
- Involved in scheduling Oozie workflow engine to run multiple HiveQL, Sqoop and Pig jobs.
- Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Pig and Hive.
- Configured Fair Scheduler to allocate resources to all the applications across the cluster.

- Performed maintenance, monitoring, deployments, and upgrades across infrastructure that supports all Hadoop clusters.
- Used Zookeeper and Oozie for coordinating the cluster and scheduling workflows.
- Managed jobs using Fair Scheduler to allocate processing resources.
- Developed job processing scripts using Oozie workflow to run multiple Spark Jobs in sequence for processing data
- Configured Zookeeper to coordinate the servers in clusters to maintain the data consistency and to monitor services.
- Automated all the jobs for pulling data from FTP server to load data into Hive tables, using Oozie workflows.
- Used Oozie workflows and coordinators for integrating MapReduce workflow- including Java REST service consumption and MongoDb/Neo4j ingress, and scheduling the data flow pipeline.

Sep 2011 **Systems Administrator**
Jan 2012 GRUPO SALINAS, Mexico

Optimized the performance of the database (DB2) of credit and collection of Elektra stores by tuning queries, creating user groups and assigning permissions and authority to them. I managed the backups and recoveries of programs and objects in the database. I automated the collection of statistics Power server performance through programs encoded in ILE-RPG and stored procedures in DB2 for i. I was a member of the Circle of Excellence of the systems management.

May 2010 **Programmer**
May 2011 Devant IT, Mexico

Developed functions, stored procedures in Transact-SQL and triggers for SQL Server 2008. I participated in the development of the system of maintenance logs to Federal Police aircraft using Java with Hibernate, Spring and ZK frameworks.

Education

NATIONAL AUTONOMOUS UNIVERSITY OF MEXICO

- Degree in Computer Engineering
- Diploma of Administration of Databases