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. 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.
* **7 years** of experience with the Hadoop ecosystem and f Data tools and frameworks.
* 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.
* Knowledge of incremental imports, partitioning, bucketing in Hive and Spark SQL 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, Sentry, Spark, Falcon, Hbase, Hive, Pig, Sentry, Ranger.
* Developed Scripts and automated data management from end to end and sync up between all the clusters.
* Strong hands on experience in Hadoop Framework and its ecosystem including but not limited to HDFS Architecture, Hive, Pig, Sqoop, HBase, MongoDB, Cassandra, Oozie, Spark RDDs, Spark DataFrames, Spark Datasets, Spark MLlib, etc.
* Worked on disaster management with Hadoop cluster.
* Involved in building a multi-tenant 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.

Technical Skills:

**Data & File Management**

*█ █ █ █ █ █ █ █ █ █*

Apache Cassandra, Apache Hbase, MapR-DB, MongoDB, Oracle, SQL Server, DB2, Sybase, RDBMS, HDFS, Parquet, Avro, JSON, Snappy, Gzip, DAS, NAS, SAN

**Project Management Methodologies**

*█ █ █ █ █ █ █ █ █ █*

Agile, Kanban, Scrum, DevOps, Lean, Six Sigma

**Cloud Services & Distributions**

*█ █ █ █ █ █ █ █ █ █*

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

**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, Apache Airflow and Camel, Apache Lucene, Kibana, X-Pack, Apache SOLR, Apache Drill, Presto, Apache Hue, Sqoop, Kibana, Tableau, AWS, Cloud Foundry

Experience:

**March 2019- Present**

**Big Data Development Engineer III/Knowledge Consultant**

CAPITAL ONE FINANCIAL CORPORATION – McLean, VA

**Description:** Founding member of four-engineer team focused on creation of targeted credit policies for Upmarket customer segment. Policies were targeted across multiple online and offline channels. Developed credit policy implementations across various Upmarket card products for approvals, credit line adjustment, and counteroffers. The counteroffer policies are estimated to approve up to 50% of Upmarket customers who otherwise would be ineligible for a card product, increasing revenues by tens of millions of dollars.

**Responsibilities:**

* Served as senior engineer in team of four engineers. Also worked with various engineering and business teams in other card segments to advise them on credit policy creation.
* Worked on-site working across several teams with senior and junior engineers
* Implemented credit policies for Upmarket card segment using Drools and Java Spring. Verified policies end-to-end in inner-sourced credit policy knowledge engine application.
* Implemented counteroffer credit policy for Upmarket card segment, estimated to approve up to 50% of customers who otherwise would be ineligible for a card product, increasing revenues by tens of millions of dollars.
* Automated end-to-end CI/CD with Jenkins pipeline across multiple environments
* Performed production support and post-release monitoring with logging and incident management software such as Splunk and PagerDuty
* Managed binary artifacts using JFrog Artifactory to ensure reliable and secure builds with convenient access to snapshot and release versions.
* Implemented Acceptance Test-driven Development (ATDD) test suites using Cucumber
* Contributed to inner-sourced cross-team applications via Enterprise GitHub pull requests, increasing stability and performance.
* Developed and distributed VBA macros for business team spreadsheets, allowing for greater automation of data entry tasks, helping reduce manual labor and errors entering the data.
* Created hands-on training labs to train other teams on how to create Upmarket credit policies for the inner-sourced knowledge engine

**Technologies:** KIE/Drools, Java Spring, Apache Maven, VBA, Cucumber, Jenkins, Amazon DynamoDB, PostgreSQL, Artifactory, Splunk, PagerDuty

**August 2018- March 2019**

**Big Data Cloud Systems Engineer III**

CAPITAL ONE FINANCIAL CORPORATION – McLean, VA

**Description:** Developed and maintained application which calculates interchange between Visa, MasterCard, and Capital One, grossing hundreds of millions of dollars in revenue. Leveraged Cloud technologies and DevOps toolchain to build scalable, performant architecture.

**Responsibilities:**

• Served as lead engineer to team of six engineers (3 senior, 3 junior). Worked on-site working across several teams with senior and junior engineers

• Developed and maintained Spark/Scala application which calculates interchange between Visa, MasterCard, and Capital One, grossing hundreds of millions of dollars in revenue.

• Created highly scalable, resilient, and performant architecture using Amazon AWS cloud technologies such as Simple Storage Service (S3), Elastic MapReduce (EMR), Elastic Cloud Compute (EC2), Elastic Container Service (ECS), Lambda, and Elastic Load Balancing (ELB).

• Built and maintained end-to-end CI/CD Jenkins pipeline across multiple environments.

• Deployed containerized applications using Docker, allowing for standardized service infrastructure.

• Monitored production software with logging, visualization, and incident management software such as Splunk, Kibana, and PagerDuty.

• Managed binary artifacts using JFrog Artifactory to ensure reliable and secure builds with convenient access to snapshot and release versions.

• Implemented Acceptance Test-driven Development (ATDD) test suites using Cucumber for Scala.

• Created infrastructure-as-code using HashiCorp Terraform to easily automate, deploy and evolve application infrastructure on demand.

• Designed process allowing business users to dynamically update card rewards. Data consumed by the interchange application, removing the need for developer intervention and reducing costly change orders.

• Created proof of concept for (and later implemented) integrating existing application architecture with inner-sourced pipeline platform, to improve architecture efficiency and ensure compliance to security standards.

• Upgraded application architecture from Spark 2.2.0 to Spark 2.4.0 to take advantage of new built-in Spark-Avro functionality.

• Created automated tools to improve the team workflow, eliminating costly and error-prone menial work.

• Led disaster recovery and incident management efforts to minimize downtime and loss of revenue.

• Maintained and updated database schemata for Total Systems (TSYS) PCI data warehouse.

• Presented training seminars on how to improve development processes and foster a best-practices work culture.

• Mentored junior engineers, providing supervision and guidance on how to improve their skills and grow their potential.

• Provided live demonstrations of software systems to non-technical, executive-level personnel, showing how the systems were meeting business goals and objectives.

• Led team daily standup and planning meetings, utilizing Agile processes such as Scrum and Kanban to direct and maximize engineer productivity.

• Published extensive team-facing and client-facing documentation for applications on Confluence.

**Technologies:** Apache Spark, Hadoop, Scala, AWS (Simple Storage Service (S3), Elastic MapReduce (EMR), Elastic Cloud Compute (EC2), Elastic Container Service (ECS), Lambda, and Elastic Load Balancing (ELB)), Java, Jenkins, Github, Jira, Confluence, Terraform, Maven, Gradle, Artifactory, Splunk, PagerDuty, Kibana

**March 2018- August 2018**

**Hadoop Enterprise Architect**

AMERICAN EXPRESS – PHOENIX, AZ

**Description:** The project was a recommender system for recommending credit products to prospective members. I was in charge of upgrading this production system and also worked to optimize maintenance and performed some administrative tasks. The system environment was a remote private cluster. We directly connected to the MapR filesystem for making changes and doing computations although we did utilize AWS for file storage and data warehousing.

**Responsibilities:**

* Worked embedded with core team on-site and collaborated remotely with Indian offshore teams.
* Oversaw Big Data engineering tasks for team of eleven on-site engineers and other offshore teams.
* Created POC for migrating existing MapR Hive on MR systems to Apache Spark 2.3.0, resulting in 1000% faster query responses.
* Trained engineering staff on engineering best practices and technologies, including TDD, BDD, automated testing, change impact analysis, Hadoop/HDFS, Apache Hive.
* Updated onboarding process and documentation for new hire engineers, shortening the onboarding process from two weeks to two days.
* Led effort in refreshing outdated automated test suite, bringing failing test count from over 150 down to zero.
* Streamlined Unix shell and Hive program scripts, decreasing program startup time by a factor of 10.
* Introduced system to create and review pull requests in Jira for feature enhancements to production software, reducing defects and regressions by more than 300%.
* Served as liaison between Big Data team, Java Development Team and Web Development teams, resolving conflicts and creating and assigning user stories in Jira.
* Set-up automated continuous integration processes using private Jenkins server.
* Standardized team project management, source control, and documentation using the Atlassian Enterprise Jira/Confluence/Bitbucket stack.
* Created UDFs and stored procedures to add customizable, modular, and reusable functionality to existing Hive, Pig, and SQL programs. The UDFs and stored procedures were written in Hive, Pig, and SQL.
* Created cron-synchronized scheduler task to automatically refresh RESTful web services cached data every 30 minutes, eliminating the need for costly and error-prone manual refreshing.
* Implemented OWASP security standards to combat common vulnerabilities, such as injection, cross-site scripting (XSS), and sensitive data exposure.

**Technologies:** MapR, Hadoop/HDFS, Hive, HBase, Spark, Pig, Java Spring, Java EE, Java ESB, REST web services, Shell, Python

**August 2017- March 2018**

**Hadoop Data Engineer / Project Manager**

CITIZEN’S BANK/NEW VANTAGE PARTNERS – BRIDGEPORT, CT

**Description:** Responsible for project management and team management. NVP embedded engineers on-site in separate teams with a handful of bank analysts. Working in parallel, we worked to take raw financial data and conform it to a dimensional model. This dimensional data would then be used to stage in IBM DataStage and load into IBM Netezza database

**Responsibilities:**

* Led a group of six business analysts, 4 analysts on-site, 2 remote analysts
* Coordinated cross-team projects and tasks with engineers and business personnel
* Applied Agile principles to team work activities
* Tracked team’s biweekly sprint assignments in Jira using Scrum and Kanban
* Supported team efforts through daily standup meetings and planning sessions, and weekly retrospective meetings.
* Trained analysts on Business-Driven Development and Test-Driven Development methodologies, including a primer on Cucumber and Gherkin.
* Created manual and automated test cases to test business workflow functionality.
* Assisted analysts in creation of SQL and HiveQL queries to test workflows and verify input and output data
* Met frequently with engineers and business personnel to make decisions on project architecture, business and technical requirements, and project scope.
* Resolved project issues stemming from missed deadlines, changing business requirements and project scope, analyst absences, and missing or incomplete data
* Trained analysts on how to migrate Podium Data workflows during upgrade from version 2 to 3.
* Discovered and documented bugs in the Podium Data platform, and worked with Podium technical representatives to fix the bugs and develop a patch for the client.
* Trained analysts on how to securely transmit and share documents in the cloud and bank intranet.
* Utilized Subversion and GitHub repositories to prevent data loss and backup critical work artifacts.
* Wrote User-defined Functions (UDFs) in order to add custom functionality to the Podium Data tool.
* Administered performance testing and recommended hardware upgrades to boost project performance and boost analyst productivity
* Wrote detailed technical documentation for project systems and supporting processes
* Coordinated work between on-site and offshore resources in order to maximize productivity and reduce analyst downtime
* Identified workflow defects and cataloged them in Jira for resolution in future sprints
* Created and presented slides to business personnel that explained complex technical details in an easy-to-understand manner
* Designed and diagrammed data models which conformed raw data to a dimensional model to facilitate ETL processes in IBM DataStage and IBM Netezza.

**Technologies:** Apache Hadoop, Apache Pig, Apache Hive, Podium Data, Java, ER Studio Data Architect, Atlassian Jira, Amazon AWS, Cucumber/Gherkin, Subversion (SVN), GitHub

**May 2016- August 2017**

**Hadoop Data Architect/Engineer**

COMCAST XFINITY – PHILADELPHIA, PA

**Description:** Involved in testing mobile wifi hotspot technology by gathering data from Comcast modems which are used for wi-fi free to subscribers, like a public utility. In exchange, Comcast gets an enormous amount of data - with the idea being that the data can be used to offset the cost of the wi-fi in other areas. For example, sensors tell Comcast about utilization, price optimization, and geographical variants.

**Responsibilities:**

* Involved in creating Hive Tables, loading with data and writing Hive queries, which will invoke and run 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 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 cluster resources 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.
* Spark Streaming with Kafka & HDFS & MongoDB to build a continuous ETL pipeline. This is used for real time analytics performed on the data.
* 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.

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

**May 2015- May 2016**

**Hadoop Data Architect/Engineer**

CITY OF LONG BEACH – LONG BEACH, CA

**Description:** 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.

**Responsibilities:**

* Analyzed Hadoop cluster using big data analytic tools including Kafka, Pig, Hive, Spark
* 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.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* 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.

**Technologies:** Hadoop, HDFS, Hive, Spark, YARN, Kafka, Pig, MongoDB, Sqoop, Storm, Cloudera, Impala

**January 2014 - May 2015**

**Hadoop Data Engineer**

GULFSTREAM – SAVANNAH, GA

**Description:** 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.

**Responsibilities:**

* 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.
* 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

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

**August 2012 - December 2013**

**Hadoop Data Engineer**

ALIBABA – REMOTE

**Description:** Alibaba works with buyers and suppliers through its web portal. Apache Storm provides it the feature to take into consideration the purchases that are being made during the day while recommending products to users. This plays a key role on special days (holidays) when the activity is unusually high. This is an example where efficient stream processing plays over batch processing

**Responsibilities:**

* Responsible for building scalable distributed data solutions using Hadoop.
* Designed and oversaw implementation of large-scale parallel relation-learning system.
* Installed and configured Pig for ETL jobs and made sure we had Pig scripts with regular expression for data cleaning.
* Creating Hive external tables to store the Pig script output. Working on them for data analysis in order to meet the business requirements.
* Developed pipeline jobs to process the data and create necessary Files.
* Involved in loading the created Files into HBase for faster access of all the products in all the stores without taking Performance hit.
* Used Zookeeper for providing coordinating services to the cluster.
* Used Oozie Scheduler system to automate the pipeline workflow and orchestrate the jobs that extract the data on a timely manner.
* Imported data using Sqoop to load data from MySQL and Oracle to HDFS on regular basis.
* Moving data from Oracle to HDFS and vice-versa using SQOOP.
* Collected and aggregated large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Used Sqoop to efficiently transfer data between databases and HDFS and used Flume to stream the log data from servers.
* Implemented partitioning, bucketing in Hive for better organization of the data.
* Worked with different file formats and compression techniques to determine standards.
* Worked on installing cluster, commissioning and decommissioning of data node, NameNode recovery, capacity planning, and slots configuration.
* Involved in loading data from Linux file system to HDFS.
* Used Linux shell scripts to automate the build process, and to perform regular jobs like file transfers between different hosts.
* Involved in production support, which involved monitoring server and error logs, and foreseeing and preventing potential issues, and escalating issue when necessary.
* Documented Technical Specs, Dataflow, Data Models and Class Models.
* Documented requirements gathered from stake holders.
* Successfully loaded files to HDFS from Teradata, and loaded from HDFS to HIVE.
* Used Zookeeper and Oozie for coordinating the cluster and scheduling workflows.

**Technologies:** Hadoop Cluster, HDFS, Hive, Pig, Sqoop, Linux, HBase, Shell Scripting, Eclipse, Oozie, Navigator.

**July 2011- July 2012**

**BI Developer**

BANK OF AMERICA – CHARLOTTE, NC

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

**Responsibilities:**

* 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.
* 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.
* Hands-on experience of large database systems: RDBMS Oracle and SQL.

**Technologies:** RDBMS, SQL, Oracle, XML.

**May 2010 - Jun 2011**

**Lead Software Engineer**

AMAZON MWS – CARROLLTON, TX

**Description:** As a sole developer, created application allowing workers to scan received product and inventory, speeding up handling and eliminating human error.

Performed Amazon inventory management, operations, and logistics.

**March 2009 - June 2010**

**Software Engineer**

ACCENTURE – IRVING, TX

**Description:** Utilized Big Data (NoSQL/Hadoop) technologies to create in-house support applications.

Provided technical support to colleagues to accomplish their project goals.