B Ravi

Senior BigData Engineer (Spark, Python/Scala, AWS)

Email: ravi.devb@gmail.com

Mobile: (484)961-0898

****

**PROFESSIONAL SUMMARY**

* Technology professional with **Java/J2EE, Spark Scala, Python, Big data applications** with 8 years of progressive and diverse experience in developing applications with an emphasis on Hadoop ecosystem Tools and Technologies using industry accepted methodologies. Strong ability and experience in working on object oriented and functional components of JVM based languages like **Java (8), Scala** and **Python 2.7 and 3.8.3**
* Experienced in Apache **Spark** for implementing advanced procedures like text analytics and processing using the in-memory computing capabilities written in Scala
* Hands on experience in working with multiple tools in the Hadoop ecosystem like Spark SQL, Spark Streaming, Hive, Pig, Sqoop, Kafka, MapReduce on various Hadoop distributions like Horton Works, Cloudera, AWS EMR Cluster
* 2 year of hands on experience with data lake implementations, core modernization and **data ingestion**
* Experience with design and development of Hadoop and Data Lake Solution components across multiple Hadoop platforms like Cloudera, Hortonworks
* Hands on experience in writing **Map Reduce** programs using Java to handle different data sets using **Map and Reduce tasks**
* Worked on front end pages that were developed in a professional manner using HTML, CSS, JavaScript and JQuery
* Knowledge of application development for cloud platforms using technologies like Java/J2EE, Spring Boot, Spring Cloud, Micro Services.
* Experience in the full life cycle of **migrating data** from existing in-house data center to the cloud built on top of **AWS** with experience in using multiple compression techniques
* Experience working with multiple RDBMS like MySQL, Oracle and columnar databases like Cassandra, Cosmos DB.
* Experience with **Devops** tools like Jenkins for continuous integration, Gitlab as a repository and chef for automation of system level patching and upgrades.
* Working with onsite and offshore team members, mentoring junior team members and ability to work in team.
* Strong debugging and critical thinking ability with fabulous understanding of frameworks advancement in methodologies and strategies.

|  |  |
| --- | --- |
| **BigData Ecosystem** | Spark Streaming, Spark SQL, Kafka, Pig, Map Reduce, Hive, Sqoop, HDFS |
| **Cloud Services** | AWS EC-2, EMR, S3, Qubole, Snowflake |
| **Languages** | Java 8, Scala, Python 2.7 and 3.8.3(Anaconda Package Manager) |
| **Scripting Language** | Unix Shell Scripting, Python Scripting |
| **NOSQL DB** | Cosmos DB, HBase, Cassandra, Elastic Stack (ELK) |
| **RDBMS DB** | Oracle, MySQL |
| **AI/ML** | Population Stability Index, H2O Machine Learning Frame Work |
| **Servers** | Tomcat, JBoss |
| **Operations** | Maven, Jenkins, SBT, GIT Lab |
| **Web Services** | REST, SOAP |
| **Markup Languages** | HTML/HTML5, XML, XML Schema, CSS/CSS3 |
| **Operating Systems** | Windows, Linux Distro (Ubuntu, Mint, Fedora) |

**Technical Skills**

**PROJECT EXPERIENCE**

**Sr Big Data Engineer (AWS, Python, Spark/Scala) Aug 2018 – Present**

**IQVIA Plymouth Meeting, PA**

IQVIA is the leading human data science company focused on helping healthcare clients find unparalleled insights and better solutions for patients. Formed through the merger of IMS Health and Quintiles, IQVIA offers a broad range of solutions that harness the power of healthcare data, domain expertise, transformative technology, and advanced analytics to drive healthcare forward.

**Responsibilities:**

* Developed Spark/Scala scripts, UDFs using both Data frames/SQL and RDD in Spark for Data Aggregation, queries and writing data back into AWS S3.
* Experience in working with Reltio API’s and consuming the data into the Spark Data frames for Performing MDM Extracts using AWS-EMR distributed Environment.
* Developed Change Data Capture (CDC) mechanism using **Spark SQL, Spark Data frames, Python, Hive** and **AWS-EC2** to capture the changes from NoSQL databases and the changes processed will be stored in **AWS-S3**
* Implemented consolidation of **CDW** module using **Spark, Scala, Python and AWS S3, EC2** by considering input file as **JSON** and the process accumulates of all monthly, weekly, daily zip files
* Built the Efficient Real-time data processing pipeline using **Kafka**, spark streaming and HBase for processing the incoming data instantly
* Written the **Kafka**-**Spark** Streaming module acting as consumer to **Kafka** which executes the business logic using spark **DStreams** and RDD methods.
* Scheduled Jobs through **Automic** by creating Work Flows, DAG’s, Breakpoints and Submit Jobs
* Worked on the Bone Patient Level Data with Stability related start date and end date for each claim. Used Scala collection framework to store and process the complex information for the **Osteoporosis** treatment
* Implemented a POC to integrate ELK Stack to an existing appliance framework for real time log aggregation, analysis and querying (Elasticsearch, Logstash, Kibana)
* Experienced in handling large datasets using **Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations**
* Experienced in Performing tuning of Spark Applications for setting right Batch Interval time, broad casting and memory tuning.
* Worked on migrating legacy Java programs into Spark transformations using **Spark** and **Scala**.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD’s Involved in every stage of the data product development life cycle from coordinating initial requirements with business owners through development with subject matter experts, to production deployments
* Used **Git lab** for version control system and code maintenance and maintained **AWS EMR and EC2** Cluster through **Qubole**

**Environment: Scala, Spark SQL, Kafka, Hive, Python, Reltio MDM, ELK, AWS EMR, S3, GitLab, Automic, Qubole, HDFS, Cassandra, DBeaver, IBM DB2**

**Sr Data Engineer (Python, Spark) April 2018 – July 2018**

**Wells Fargo Concord, CA**

Monitoring is the process of paying constant attention to the accuracy of the models to decide when a machine learning model should be re-calibrated or rebuild. Managing and monitoring ML models is challenging when there are many models used by multiple people over an extended time. However, the risks of not managing and monitoring them effectively can be catastrophic, as it allows machine learning models to become outdated, leading to potential erroneous and costly business decisions. This is because the behavior of data changes over time, often in the ways that the machine learning model wasn’t built to anticipate. Effective model management and monitoring detects when the data behavior and the model must be re-calibrated

**Responsibilities:**

* Population Stability Index (PSI) is a metric to measure how much a variable has shifted in distribution between two samples or over time.
* Developed pyspark applications to generate PSI values, Histograms, VBW Histograms and QQ Plots.
* Experience with Pandas Data Frames, Spark Data Frames and Data Sets.
* Have good experience with real time streaming messaging systems like Kafka.
* Calculated percentage of records in each group based on scoring samples and training samples
* Reading Parquet files from Hive Data lake and summarizing the data by using Spark transformations and storing the resulted data on a NoSql databases in a json format
* Developing a Rest API calls to fetch the data and represent them on UI Dash board
* Used AutoSys scripts to automate the workflow
* Experience in drill connectivity to fetch the data from NoSql database
* Providing inputs to team managers and other researches to help define research vision and informing technical direction
* Building Proof-Of-Concepts around projects that can demonstrate utilization, value and lead to scalable solutions

**Environment: Python/PySpark, Spark SQL, Spark Streaming, Kafka, AutoSys, Rest API, Redis Cache, Maven, Datalake, Jenkins, Model Monitoring**

**Sr Big Data Consultant (Spark, Scala, Azure) Oct 2017 – March 2018**

**Johnson Controls International Milwaukee, WI**

The IoT Platform team in the Data Enabled Business (DEB) at JCI builds, delivers, maintains and supports a suite of micro-services that enable sending data from sensors on JCI equipment at customer sites to the Cloud, storing the data securely on the cloud, analyzing this data in real-time as it streams in and providing metrics based on these analysis. These micro-services are used by various JCI Connected Applications to deliver “Big Data And cloud-enabled” features such as Advanced Analytics, Data Virtualization and Command & Control within their applications.

**Responsibilities:**

* Hands on experience in developing Batch Processing pipeline from end to end by using spark data frames and Scala
* Migrated Billions of Historic Data into Azure CosmosDB by developing applications using Spark
* Populating the Rest API Data into many fields as per the requirement and sending the batches of data to the Kafka and storing the result into **Hbase**
* Worked on HBase in creating **HBase** tables to load large sets of semi structured data coming from various sources.
* Created data science workflows using Packer, Terraform, and Zeppelin which greatly increased productivity.
* Experienced import/export data into **Hive** from relational data base and **Tera Data** using **Sqoop**.
* Developed Spark application for Generating UUIDs for the Chillers, registering them and setting up the permissions through Rest API Services by using spark
* Hands on experience on creating Docker images and running it on Kubernetes Cluster
* Load the data into Spark RDD and performed in-memory data computation to generate the output response.
* Performed different types of transformations and actions on the RDD to meet the business requirements.
* Developed and configured Kafka to pipeline server logs data into spark streaming for real time processing
* Developed Micro Services using Spring MVC, Spring Boot, Spring Cloud.
* Used a Microservices architecture with Spring Boot-based services interacting through a combination of REST and Spring Boot.
* Used Microservices designed with the individual database and project with no dependencies.
* Experience in Elastic Search in Hadoop that bridges that gap, letting us the leverage for the best of Hadoop's big data analytics and the real-time analytics
* Enabled speedy reviews and first mover advantages by using Apache Oozie workflow engine for managing and scheduling Hadoop jobs
* Developed MapReduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables
* Familiar with data architecture including data ingestion pipeline design, Hadoop information architecture, data modeling and data mining, machine learning and advanced data processing

**Environment: Scala, Spark SQL, Azure, Apache Zeppelin, Spark Streaming, Kafka, Oozie, Rest API, Redis Cache, Maven, Jenkins, Kubernetes**

**Sr Big Data Consultant (Spark, Kafka) Feb 2016 – Sep 2017**

**Global Pharma Tek Edison, New Jersey**

Data Lake is a centralized repository for storing and managing the data across the enterprise which is being developed for a client who has sponsored the project. The framework is primarily used to acquire, ingest and provision the data. The stored data is used for further analytics, which are specific to CRO/Pharma/Biotech business using the analytical frameworks.

**Responsibilities:**

* Actively participated from design phase of the data lake starting with performing POC’s using multiple Big Data **(Cassandra, HBase, Kafka, DynamoDB, AWS)** tools to identify the best tools that solve the business problem at hand.
* Shared responsibility for the complete life cycle of migrating data from the existing in-house infrastructure to the cloud built on top of **AWS**-**S3**.
* Developed **Impala** to define the tables and mapping them to equivalent tables in HBase.
* Involved in the development of **Spark** Streaming Application for one of the data sources using **Scala, Spark** by applying the transformations
* Involved in converting Hive/SQL queries into **Spark transformations** using **Spark RDD**, **Scala** and **Python**
* Developed a Script in **Scala** to read all the Parquet tables in a Database and parse them as Json files another script to parse them as a structured table in Hive.
* Extensively involved in writing Spark Applicationswith **Spark-SQL/Streaming** for faster processing of Data.
* Analyzed schedules in **Control-M** to check for dependencies, modified to add dependencies on new jobs using xmls
* Responsible for design and creation of **Hive tables, partitioning, bucketing, loading data** and **writing hive queries.** Implemented, migrated existing Hive Script in **SparkSQL** for better performance.
* Created **RESTful** services, converted data formats to make it consumable by services.
* Developed **Spark streaming** jobs in **JAVA 8** to receive real time data from **Kafka,** process and store the data to HDFS.
* Experience in validating and cleansing the data using **Pig** statements and hands-on experience in developing Pig **MACROS**.
* Used **Agile (SCRUM)** methodologies for Software Development.

**Environment: JAVA8, Sqoop, Impala, Kerberos, Spark SQL, Spark Streaming, Kafka, Scala, Amazon Web Services (AWS-EMR), Hive, Pig, REST, Oozie, Control-M, Maven, Jenkins.**

**Big Data Engineer April 2015 – Nov 2015**

**VAQYA, (HowUSayIt.Com) New Jersey**

Vaqya is a voice-enabled social networking platform that will enable users to create, share and follow users and content a multi-modal interface including voice, touch and text.

**Responsibilities:**

* Actively participated in complete software development lifecycle (Scope, Design, Implement, Deploy, Test) including design and code reviews.
* Moving Bulk amount of data into **HBase** with **Map Reduce Integration.**
* Developed Pig Programs for loading and filtering the streaming data into HDFS usin**g Kafka**.
* Developed Scala Scripts, UDFs using both **Data frames/SQL and RDD/MapReduce** in **Spark 1.6** for Data Aggregation, Queries and writing data back into OLTP System through Sqoop.
* Performed advanced procedures like text analytics and processing, using in-memory computing capabilities of **Spark** using **Scala**
* Developed **HBase data model** on top of HDFS data to perform real time analytics.
* Used tools like **MapReduce and Spark with Scala** for performing operations like Clickstream Analysis and to perform Analysis on batch Data.
* Experienced in handling Avro data files by passing schema into HDFS using **Avro tools and Map Reduce.**
* Optimizing the **Hive** queries using **Partitioning** and **Bucketing** techniques, for controlling the data distribution.
* Developed visualizations and dashboards using multiple BI tools like **Tableau, Platfora**.
* Used **Oozie** scheduler to automate the pipeline workflow and orchestrate the **Map Reduce**, **Sqoop**, **hive** and **pig** jobs that extract the data on a timely manner.
* Involved in story-driven agile development methodology and actively participated in daily scrum meetings.

**Environment: MapReduce, Spark with Scala, Hive, Pig, Sqoop, Oozie, HBase, Platfora, Redis, REST Services, Linux, Maven, Jenkins, HDFS.**

**Java/J2EE Developer Nov 2011 - July 2014 Peritus Technologies Hyderabad, India**

Peritus Technologies is developing a product that supports their executives in processing and maintaining billing details of their customers.

**Responsibilities:**

* Designed, developed and validated User Interface using HTML, Java Script, XML.
* Handled the database access by implementing Controller Servlet.
* Implemented PL/SQL stored procedures and triggers.
* Extensively worked with Struts, Hibernate, Spring (Spring Core, Spring MVC) application design, development.
* Experience in using various design patterns like Business Delegate, Session Facade, Service Locator, Singleton and Model-View-Controller
* Extensively worked with Web Services including SOAP over HTTP, REST
* Experience with multithreading, enterprise java beans, worked with session, entity and message driven bean
* Experience in preparation of Test procedures, Test Scenarios, Test Cases and Test Data.
* Created Test Cases using Element Locators and Selenium WebDriver Methods.
* Execution of Selenium Test Cases and Reporting Defects.
* Involved in Regression Testing and Automation Infrastructure Development using Selenium.
* Expertise in implementation of Automation FrameWork using Selenium, LoadRunner, UFT.
* Experience in developing applications in WebSphere 8.5 & 7, JBoss, Tomcat and BEA web logic. Developed server running script for automation using the JBoss 6.3 application server.
* Strong experience in SOA(Service Oriented Architecture), EAI (Enterprise Application Integration) and ESB (Enterprise Service Bus)
* Extensive experience in developing web page quickly and effectively using JSF, Ajax JQuery, JavaScript, HTML, CSS and also in making web pages cross browser compatible
* Good in unit testing skills using Junit framework and functional Junit capturing user entered data and mapping it back to database to provide accurate test results.
* Interacted with application architect to design the workflow and service integration on top of spring MVC, Ajax and web services layers.
* Set up Web sphere Application server and used Ant tool to build the application and deploy the application in Web sphere
* Used Spring Framework for Dependency Injection and integrated with Hibernate.
* Involved in writing JUnit Test Cases.
* Used Log4J for any errors in the application

**Environment: Java, Spring Framework, J2EE, HTML, JUnit, XML, JavaScript, Eclipse, WebLogic, PL/SQL, Maven, Oracle.**

**EDUCATION**

**Texas A&M University JNTU, India**

MS in Computer Science Engineering Bachelors in Information Technology

GPA: 3.9/4.0 GPA: 3.6/4.0

**CERTIFICATIONS**

Horton Works Certified Professional

**REFERENCES:** Shall be provided on request