

Bapuji  
Sr. Hadoop Developer  
Phone: +1(224)-706-0020  
Email: [bapuji.dbj@gmail.com](mailto:bapuji.dbj@gmail.com)

---

### **PROFESSIONAL SUMMARY:**

- Over **8+** years of experience including **4** years of Big Data Ecosystem related technologies with full project development, implementation and deployment.
- Strong Experience working with various Hadoop ecosystem components like, **Map Reduce, HDFS, Hive, Sqoop, Pig, Flume, and Oozie**.
- Strong Knowledge on **Architecture** of **Distributed** systems and **Parallel processing** frameworks.
- In-depth understanding of **MapReduce** Framework and **Spark** execution model.
- Worked extensively on fine-tuning long running **Spark** Applications to utilize better **parallelism** and **executor** memory for more caching.
- Strong experience working with both **batch** and **real-time** processing using **Spark** framework.
- Expertise in developing production ready **Spark** applications utilizing **Spark-Core, Data frames, Spark-SQL, Spark-ML** and **Spark-Streaming** API's.
- Hands on experience in installing, configuring and deploying **Hadoop** distributions in cloud environments (**Amazon Web Services**).
- Expertise in developing production ready **Spark** applications utilizing **Spark-Core, Data frames, Spark-SQL, Spark-ML** and **Spark-Streaming** API's.
- Worked on building real time data workflows using **Kafka, Spark streaming** and **HBase**.
- Worked extensively on **Hive** for building complex data analytical applications.
- Very good understanding of **Partitions, bucketing** concepts in **Hive** and designed both **Managed** and **External** tables in Hive to optimize performance.
- Used custom serDes like **Regex** SerDe, **JSON** SerDe, **CSV** SerDe etc., in hive to handle multiple formats of data.
- Having knowledge in **Apache Ambari** platform for securing, managing and monitoring Hadoop clusters.
- Experienced in Cluster coordination services through **zookeeper**.
- Strong experience using different columnar file formats like **Avro, RCFile, ORC** and **Parquet** formats.
- Worked with **Sqoop** to move (import/export) data from a relational database into Hadoop.
- Experience working with Hadoop clusters using **Cloudera, Amazon EMR** and **Hortonworks** distributions.
- Extensive experience in performing **ETL** on structured, semi-structured data using **Pig** Latin Scripts.
- Designed and implemented **Hive** and **Pig UDF's** using **Java** for evaluation, filtering, loading and storing of data.
- Experienced in job workflow scheduling and monitoring tools like **Oozie**.
- Well versed with **UNIX** and **Linux** command line and **shell** script.
- Adequate knowledge and working experience with **agile** methodology.

### **TECHNICAL SKILLS:**

<b>Languages</b>	Java, Scala, SQL, PL/SQL, Pig Latin, Python, Hive QL
<b>Web Technologies</b>	JEE (JDBC, JSP, SERVLET, JSF, JSTL), AJAX, JavaScript
<b>Big Data Systems</b>	Hadoop, HDFS, MapReduce, YARN, Pig, Hive, Sqoop, Flume, Oozie, Impala, Spark, Apache Airflow, Kafka, Splunk, Cloudera CDH4, CDH5, Hortonworks, Hadoop EMR, Talend and Ranger...
<b>RDBMS</b>	Oracle 10g/11g, MySQL, SQL Server 2005/2008 R2, PostgreSQL, DB2, Teradata
<b>NoSQL Databases</b>	HBase, MongoDB, Cassandra
<b>App/Web Servers</b>	Apache Tomcat, WebLogic

<b>SOA</b>	Web services, SOAP, REST
<b>Frameworks</b>	Struts 2, Hibernate, Spring 3.x
<b>Version Control</b>	GIT, CVS, SVN
<b>IDEs</b>	Eclipse, Scala IDE, NetBeans, IntelliJ IDEA
<b>Operating Systems</b>	UNIX, Linux, Windows

## **EDUCATION:**

**Bachelor of Technology** in **Computer Science Engineering** at **JNTU, Kakinada**, Andhra Pradesh, India.

## **WORK EXPERIENCE:**

**Cigna – Bloomfield, Connecticut**

**Jul'17 – Present**

**Role: Hadoop/Spark Developer**

### **Responsibilities:**

- Developed **Spark** applications using **Scala** utilizing **Data frames** and **Spark SQL** API for faster processing of data.
- Developed highly optimized **Spark** applications to perform various data **cleansing, validation, transformation** and **summarization** activities according to the requirement
- Data pipeline consists **Spark, Hive** and **Sqoop** and **custom built Input Adapters** to ingest, transform and analyze operational data.
- Developed **Spark** jobs and **Hive** Jobs to summarize and transform data.
- Used **Spark** for interactive queries, processing of **streaming** data and integration with **NoSQL** database **HBase, Cassandra** for interactive access patterns.
- Involved in converting **Hive** queries into **Spark** transformations using **Spark Data Frames** in **Scala**.
- Automated creation and termination of **AWS EMR** clusters using **AWS, java sdk**.
- Built **real time** data pipelines by developing **Kafka** producers and **spark streaming** applications for consuming.
- Ingested **syslog** messages to **Kafka**.
- Worked on **Apache Airflow** to schedule single and sometimes complex chains of tasks that depend on each other on regular intervals.
- Handled importing data from relational databases into **HDFS** using **Sqoop** and performing transformations using **Hive** and **Spark**.
- Having knowledge in **Apache Ambari** platform for securing, managing and monitoring Hadoop clusters.
- Exported the processed data to the relational databases using **Sqoop**, to further visualize and generate reports for the BI team.
- Experienced in cluster coordination services through **Zookeeper**.
- Installed, tested and deployed monitoring solutions with **Splunk** services.
- Used **Hive** to analyze the **partitioned** and **bucketed** data and computed various metrics for reporting.
- Developed **Hive** scripts in **Hive QL** to de-normalize and aggregate the data.
- Scheduled and executed workflows in **Oozie** to run various jobs.
- Designing & creating ETL jobs through **Talend** to load huge volumes of data into Cassandra, Hadoop Ecosystem and relational databases.

**Environment:** Hadoop, Spark, Hive, Java, Scala, Maven, Impala, Oozie, Oracle, Ambari, GitHub, Tableau, Unix, Hortonworks, Apache Airflow Kafka, Zookeeper, Sqoop, Cassandra, Talend, Splunk, HBase.

**Qualcomm -- San Diego, CA**

**Dec'16 – Jun'17**

**Role: Hadoop/Spark Developer**

## Responsibilities:

- Part of **Big Data Center of Excellence (CoE)**, responsible for designing and building enterprise data analytics platform.
- Worked with respective business units in understanding the scope of the analytics requirements.
- Performed core **ETL** transformations in **Spark**.
- Automated data pipelines which involve data **ingestion**, data **cleansing**, data **preparation** and data **analytics**.
- Created end to end **Spark** applications using **Scala** to perform various data **cleansing**, **validation**, **transformation** and **summarization** activities on **user behavioral** data.
- Developed end-to-end data pipeline using **FTP Adaptor**, **Spark**, **Hive** and **Impala**.
- Implemented **Spark** utilizing **Spark-SQL** heavily for faster development, and processing of data.
- Exploring with **Spark** for improving the performance and optimization of the existing jobs in Hadoop using **Spark-SQL**, Data Frame running in **Yarn** mode.
- Handled importing other enterprise data from different data sources into **HDFS** using **Sqoop** and performing transformations using **Hive**, **Map Reduce** and then loading data into **HBase** tables.
- Collecting and aggregating large amounts of log data using **Flume** and staging data in **HDFS** for further analysis
- Wrapper developed in **Python** for instantiating multithreaded application and running with other applications.
- Analyzed the data by performing Hive queries (**Hive QL**) and running Pig scripts (**Pig Latin**) to study customer behavior.
- **Data warehousing**, experience in design, development and testing, implementation and support of enterprise **data warehouse**.
- Used **Hive** to analyze the partitioned and bucketed data and compute various metrics for reporting.
- Created components like **Hive UDFs** for missing functionality in **HIVE** for analytics.
- Worked on various performance optimizations like using **distributed cache** for small datasets, **Partition**, **Bucketing** in **Hive** and **Map Side joins**.
- Created **Oozie** workflows and coordinators to automate data pipelines daily, weekly and monthly.
- Automated creation and termination of **AWS EMR** clusters using **AWS**, java sdk.

**Environment:** AWS EMR, Hadoop, Spark, Hive, Sqoop, HBase, UNIX, Talend, Pig, Linux, Java, Scala, Python, Ambari, Zookeeper.  
Hortonworks

**McKesson - Alpharetta, GA**

**Dec'15 – Nov'16**

**Hadoop/Spark Developer**

## Responsibilities:

- Developed multithreaded **Java** based Input adaptors for ingesting **click stream data** from external sources like **ftp server** and **S3** buckets on daily basis.
- Created various **spark** applications using **Scala** to perform various enrichment of these click stream data combined with enterprise data of the users.
- Implemented batch processing of jobs using **Spark Scala API**.
- Developed **Sqoop** scripts to import/export data from **Oracle** to **HDFS** and into **Hive** tables.
- Stored the data in **columnar** formats using **Hive**.
- Involved building and managing **NoSQL** Database models using **HBase**.
- Worked in **Spark** to read the data from **Hive** and write it to **Hbase**.
- Optimized the **Hive** tables using optimization techniques like **partitions** and **bucketing** to provide better performance with **Hive QL** queries.
- Worked with multiple file formats like **Avro**, **Sequence**, **Parquet** and **Orc**.

- Converted existing **MapReduce** programs to **Spark** Applications for handling semi structured data like **JSON** files, **Apache** Log files, and other custom log data.
- Loaded the final processed data to **HBase** tables to allow downstream application team to build rich and data driven applications.
- Worked with a team to improve the performance and optimization of the existing algorithms in Hadoop using **Spark**, **Spark -SQL**, Data Frame.
- Implemented business logic in **Hive** and written **UDF**'s to process the data for analysis.
- Used **Oozie** to define a workflow to coordinate the execution of **Spark**, **Hive** and **Sqoop** jobs.
- Addressing the issues occurring due to the huge volume of data and transitions.
- Designed, documented operational problems by following standards and procedures using **JIRA**.

**Environment:** Java, Hadoop 2.1.0, Map Reduce2, Spark, Unix, Pig 0.12.0, Hive 0.13.0, Linux, Sqoop 1.4.2, Flume 1.3.1, Eclipse, AWS EC2, and Cloudera CDH 4.

**American Home Shield - Memphis, TN**

**Dec'14 – Nov'15**

**Role: Hadoop Developer**

#### **Responsibilities:**

- Migrated the needed data from **MySQL** into **HDFS** using **Sqoop** and importing various formats of flat files in to **HDFS**.
- Mainly worked on **Hive** queries to categorize data of different claims.
- Involved in loading data from **LINUX** file system to **HDFS**
- Written customized **Hive** UDFs in **Java** where the functionality is too complex.
- Implemented **Partitioning**, Dynamic Partitions, Buckets in **HIVE**.
- Designing and creating **Hive** external tables using shared meta-store instead of derby with partitioning, **dynamic partitioning** and **buckets**.
- Generate final reporting data using Tableau for testing by connecting to the corresponding **Hive tables** using **Hive ODBC** connector.
- Responsible to manage the test data coming from different sources
- Reviewing peer table creation in **Hive**, data loading and queries.
- Weekly meetings with technical collaborators and active participation in code review sessions with senior and junior developers.
- Monitored System health and logs and respond accordingly to any warning or failure conditions.
- Gained experience in managing and reviewing **Hadoop** log files.
- Involved in scheduling **Oozie** workflow engine to run multiple **Hive** and **pig** jobs
- Involved **unit testing**, interface testing, system testing and user acceptance testing of the workflow tool.
- Created and maintained Technical documentation for launching **Hadoop Clusters** and for executing **Hive** queries and **Pig** Scripts

**Environment:** Apache Hadoop, HDFS, Hive, Map Reduce, Core Java, Pig, Sqoop, Cloudera CDH4, Oracle, MySQL.

**Protective Life - Edina, MN**

**Oct'13 - Nov'14**

**Role: Java Developer**

#### **Responsibilities:**

- Implemented a Web based Application using Servlets, **JSP**, spring, **JDBC**, **XML**.
- Involved in writing Spring Configuration **XML** file that contains declarations and other dependent objects declarations.
- Used hibernate to connect to Database to create the **DAO** layer.

- Developed Application Framework using Model-View-Controller using the technology Spring.
- Used **HTML**, **XHTML**, **XML**, **XSLT**, **XPATH**, **JSP** and Tag Libraries to develop view pages
- Multilayer Applications construction using Open **JPA**, **HTML5**, **Spring MVC**.
- Annotated **Spring** Architecture (Spring Beans)
- Implemented **UNIX shell** scripts to migrate various data files to S&P ratings repository
- Implemented smooth pagination capability using **JSP** to remove existing pagination utility
- Worked on **Geo API** to provide geological access capability to S&P.com site.
- Involved in **Agile** process to streamline development process with iterative development.
- Code reviews and Managing the **CVS** Repository.
- Prepare builds for **DEV** and **UAT** environments.
- Participating in the regular team meetings sprint planning meetings, user story review meetings etc.
- Involved in preparing High & low level design docs with **UML** diagrams using Microsoft **VISIO** tool.

**Environment:** JDK 1.5, XML, HTML, XHTML, JSP, Spring, DAO, Oracle Express edition, Apache ANT, CVS, Junit, UNIX, Log4J, CSS Style Sheets, Apache Tomcat, J2EE, Maven 3

**Accenture – Hyderabad, India**  
**Role: Java Developer**

**Oct'11– Sep'13**

**Responsibilities:**

- Involved in Requirements **analysis**, **design**, and **development** and **testing**.
- Involved in setting up the different roles & maintained **authentication** to the application.
- Designed, deployed and tested **Multi-tier application** using the Java technologies.
- Involved in front end development using **JSP**, **HTML** & **CSS**.
- Implemented the Application using **Servlets**
- Deployed the application on **Oracle** Web logic server
- Implemented **Multithreading** concepts in java classes to avoid deadlocking.
- Used **MySQL** database to store data and execute **SQL** queries on the backend.
- Prepared and Maintained **test environment**.
- Tested the application before going live to production.
- Documented and communicated **test result** to the team lead on daily basis.
- Involved in weekly meeting with team leads and manager to discuss the issues and status of the projects.

**Environment:** J2EE (Java, JSP, JDBC, Multi-Threading), HTML, Oracle Web logic server, Eclipse, MySQL, JUnit.

**Golan Technologies – Hyderabad, India**  
**Role: Java Developer**

**Jun'09 - Sep'11**

Golan Technologies range from turnkey solutions to custom, client-driven solutions in a variety of product categories including website development and platform based applications, demand intelligence and business insight generation. Smart sites have the ability to provide a unified user experience and consistent messaging on websites across the globe, driving a favorable brand impression.

**Responsibilities:**

- Involved in the analysis, design, implementation, and testing of the project.
- Developed UI using **HTML**, **JavaScript**, **CSS** and **JSP** for interactive cross browser functionality and complex user interface.
- Implemented the end-to-end functionality of the client requirement during the development phase.
- Implemented the functionality of **mapping** entities to the database using **Hibernate**.

- Written **SQL queries** involved in the **JDBC** connection in accordance with the business logic.
- Performed various levels of **unit testing** for the entire application using the **test cases**, which included preparation of detail documentation for the results.
- Actively participated in client meetings and taking the inputs for the additional functionality.
- Involved in fixing bugs and **unit testing** with test cases using **JUnit**.

**Environment:** J2EE, Spring, Hibernate, JavaScript, CSS, Servlets, MySQL