# prem wilfred

Sr Hadoop Developer

(917)858-2819

[premhwp@gmail.com](mailto:premhwp@gmail.com)

# summary

* Hadoop Developer with around **8 years of experience in Information Technology & 4+ years in Hadoop Ecosystem.**
* Expertise in Hadoop Ecosystem **components HDFS, Map Reduce, Hive, Pig, Sqoop, Hbase and Flume** for Data Analytics.
* Have a hands-on experience on fetching the live stream data from DB2 to Hbase table using **Spark Streaming** and Apache Kafka.
* Capable of processing large sets of structured, semi-structured and unstructured data sets.
* Experience in job workflow scheduling and monitoring tools like **Oozie and Zookeeper**.
* Expertise in writing Map-Reduce Jobs in Java for processing large sets of structured, semi-structured and unstructured data sets and store them in HDFS.
* Experience in developing Custom UDFs for datasets in Pig and Hive.
* Proficient in designing and querying the NoSQL databases like HBase.
* Experience in designing and developing tables in HBase and storing aggregated data from Hive Table.
* Knowledge on integrating different eco-systems like HBase - Hive, HBase - Pig
* Experience on streaming data using Apache Flume.
* Good Knowledge in **Apache Spark and SparkSQL**.
* Skilled on migrating the data from different databases to Hadoop HDFS and Hive using Sqoop.
* Deep Knowledge in the core concepts of **MapReduce Framework and Hadoop ecosystem**
* Analyzed large structured datasets using Hive's data warehousing infrastructure
* Extensive knowledge of creating manage tables and external tables in Hive Eco system.
* Worked extensively in design and development of business process using SQOOP, PIG, HIVE, HBASE
* Knowledge on Spark framework for batch and real time data processing.
* Knowledge on Scala Programming Language.
* Hands on experience in database design using **PL/SQL** to write **Stored Procedures, Functions, Triggers** and strong experience in writing complex queries, using **Oracle, DB2,** and **MySQL**.
* Hands on Experience in Designing and Executing of Test Cases to ensure application functionality using Selenium IDE, Selenium Web driver, Firebug and TestNG.
* Good knowledge in Software Development Life Cycle (SDLC) and Software Testing Life Cycle (STLC).

# education

Bachelor of Technology – Computer Science Engineering, JNTU, Hyderabad, India.

# TECHNICAL SKILLS

|  |  |
| --- | --- |
| **Big Data Frameworks** | Hadoop, Spark, Scala, Hive, Kafka, AWS, Cassandra, HBase, Flume, Pig, Sqoop, Map Reduce, Cloudera, Mongo DB. |
| **Big data distribution** | Cloudera, Amazon EMR |
| **Programming languages** | Core Java, Scala, Python, SQL, Shell Scripting |
| **Operating Systems** | Windows, Linux (Ubuntu) |
| **Databases** | Oracle, SQL Server |
| **Designing Tools** | Eclipse |
| **Java Technologies** | JSP, Servlets, Junit, Spring, Hibernate |
| **Web Technologies** | XML, HTML, JavaScript, JVM, JQuery, JSON |
| **Linux Experience** | System Administration Tools, Puppet, Apache |
| **Web Services** | Web Service (RESTful and SOAP) |
| **Frame Works** | Jakarta Struts 1.x, Spring 2.x |
| **Development methodologies** | Agile, Waterfall |
| **Logging Tools** | Log4j |
| **Application / Web Servers** | Apache Tomcat, WebSphere, Weblogic |
| **Messaging Services** | ActiveMQ, Kafka, JMS |
| **Version Tools** | Git, SVN and CVS |
| **Analytics** | Tableau, SPSS, SAS EM and SAS JMP |

# Professional Experience

**Client: Walgreens, Deerfield, IL**

**Hadoop Developer**, Jan 2017 – present

**Responsibilities:**

* Maintained System integrity of all sub-components (primarily **HDFS**, **MR**, **HBase**, and **Hive**).
* Integrated the **hive** warehouse with **HBase**
* Migrating the needed data from MySQL into HDFS using **Sqoop** and importing various formats of flat files into **HDFS.**
* Load the data into HBase tables for UI web application.
* Written customized **Hive UDFs** in Java where the functionality is too complex.
* Maintain System integrity of all sub-components related to Hadoop.
* Designed and created Hive external tables using shared meta-store instead of derby with partitioning, dynamic partitioning and buckets.
* **HiveQL** scripts to create, load, and query tables in a Hive.
* Supported **Map Reduce** Programs those are running on the cluster
* Monitored System health and logs and respond accordingly to any warning or failure conditions.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.
* Worked on Big Data Integration and Analytics based on **Hadoop, SOLR, Spark, Kafka, Storm and web Methods technologies**.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
* Real time streaming the data using Spark with Kafka.
* Worked on migrating MapReduce programs into Spark transformations using Spark and Scala
* Generate final reporting data using **Tableau** for testing by connecting to the corresponding Hive tables using Hive ODBC connector.
* Strongly recommended to bring in **Elastic Search** and was responsible for installing, configuring and administration.
* Developing and maintaining efficient ETL Talend jobs for Data Ingest.
* Worked on Talend RTX ETL tool, develop jobs and scheduled jobs in Talend integration suite.
* Modified reports and Talend ETL jobs based on the feedback from QA testers and Users in development and staging environments.
* Involved in migration Hadoop jobs into higher environments like SIT, UAT and Prod.

**Environment**: Hortonworks Hadoop2.3, HDFS, Hive, Scala, Map Reduce, Storm, Java, HBase, Pig, Sqoop, Shell Scripts, Oozie, Coordinator, MySQL, Tableau, Elastic search, Talend and SFTP.

**Client: United Airlines, Chicago IL**

**Hadoop Developer Jan 2016 – Dec 2016**

**Responsibilities:**

* Experience in supporting and managing **Hadoop Clusters** using **Hortonworks** distributions by deploying it on AWS cloud.
* Collected aggregated large amount of web log data from different sources such as web servers, mobile and network devices using **Apache Kafka**.
* Ingestion framework was developed in python.
* Creating the **RDD’s, Data frames** for faster execution and performing data transformations and actions using **Spark**.
* Developed optimal strategies for distributing the web log data over the cluster.
* Implemented **Hive Generic UDF's** to incorporate business logic into **Hive Queries**.
* Configuring **Spark Streaming** to receive real time data from the **Kafka** for high speed data processing and Store the **stream data** to HDFS.
* Used Scala to read text data, CSV data, image data from HDFS, S3 and Hive
* Worked on **Spark SQL** for faster execution of Hive queries using **Spark SQL Context**.
* Invoking **HQL** from **Spark SQL** and storing it in **parquet file** as the storage format.
* Creating **Hive tables** and working on them using **HQL** for evaluation, filtering, loading and storing of data.
* Design and implemented Incremental Imports using **Sqoop** into Hive tables.
* Analyzed the web log data using the **HQL** to extract number of unique visitors per day, page views, visit duration, most visited page on website.
* Developed **ETL** jobs and scheduling it in **Oozie** and performing ingestion into HDFS.
* Cluster authentication(security) is maintained by using **Kerberos**.
* Good experiences in designing and developing various kinds of data visualizations, dashboards, stories, and workbooks in **Tableau**.
* Involved in Agile methodologies, daily scrum meetings, sprint planning.

**Environment:** HDFS, Hive, Sqoop, Oozie, Storm, Scala 2.11.8, Spark 2.0, Spark SQL, Spark streaming, Python, Kafka, GitHub, Hortonworks(HDP), Kerberos, AWS, Amazon S3, Amazon EC2, Amazon EBS, Tableau.

**Client: Intermountain Healthcare (IHC), Salt Lake City, Utah**

**Hadoop Developer**, Oct 2014 – Dec 2015

**Description:** Intermountain is the most diversified health care company in the United States and a leader worldwide in helping people live healthier lives and helping to make the health system work better for everyone. The particular project involves in assigning a personal physician for each and every customer respective to particular area. Categorizing the diseases encountered frequently based on the claims and on the geographical location. The particular project also analyses the effectiveness of campaigns conducted at various places.

**Responsibilities:**

* Developed Big Data Solutions that enabled the business and technology teams to make data-driven decisions on the best ways to acquire customers and provide them business solutions.
* Installed and configured Apache Hadoop, Hive, and HBase.
* Worked on Hortonworks cluster, which was used to process the big data.
* Developed multiple map reduce jobs in java for data cleaning and pre-processing.
* Sqoop was used to pull data into Hadoop distributed file system from RDBMS and vice versa
* Defined workflows using Oozie.
* Used Hive to create partitions on hive tables and analyzes this data to compute various metrics for reporting.
* Created Data model for Hive tables
* Good Experience in managing and reviewing Hadoop log files
* Used Pig as ETL tool to do transformations, joins and pre-aggregations before loading data onto HDFS.
* Worked on large sets of structured, semi structured and unstructured data
* Responsible to manage data coming from different source
* Installed and configured Hive and developed Hive UDFs to extend core functionality of hive
* Responsible for loading data from UNIX file systems to HDFS.
* Arrange for on-site job coaching or assistive devices, such as specially equipped wheelchairs, to help clients adapt to work or school environments.
* Locate barriers to client employment, such as inaccessible work sites, inflexible schedules, and transportation problems, and work with clients to develop strategies for overcoming these barriers.
* **Environment:** Apache Hadoop 2.x, MapReduce, HDFS, Hive, HBase, Pig, Oozie, Linux, Java 7, Eclipse Mars with physicians, psychologists, occupational therapists, and other professionals to develop and implement client rehabilitation programs.

**Optum Insight, Inc. Eden Prairie, MN**

**Hadoop Developer-Health Care,** Aug 2013 – Oct 2014

Optum Insight provides software and information products, advisory consulting services and business process outsourcing to participants in the health care industry.

**Responsibilities: -**

* Implemented Big Data platforms using Cloudera CDH4 as data storage, retrieval and processing systems.
* Involved in installing, configuring and managing Hadoop Ecosystem components like Hive, Pig, Sqoop and Flume.
* Responsible for loading data from Teradata database into a Hadoop Hive data warehousing layer, and performing data transformations using Hive.
* Partitioned the collected people's data by disease type and medication prescribed to the patient to improve the query performance.
* Used Hive Query language (HQL)  to further analyze the data to identify issues and behavioral patterns.
* Used Apache Spark and Scala language to find patients with similar symptoms in the past and medications used for them to achieve best results.
* Configured Apache Sentry server to provide authentication to various users and provide authorization for accessing services that they were configured to use.
* Worked on Kafka while dealing with raw data, by transforming into new Kafka topics for further consumption.
* Installed and configured Hortonworks Sandbox as part of POC involving Kafka-Storm-HDFS data flow.
* Log data from webservers across the environments is pushed into associated Kafka topic partitions, SparkSQL is used to calculate the most prevalent diseases in each city from this data.
* Implemented Zookeeper for job synchronization
* Deployed NoSQL database HBase to store the outputs of the jobs.
* Supported operations team in Hadoop cluster maintenance activities including commissioning and decommissioning nodes and upgrades.
* Extended the core functionality of Hive language by writing UDF's and UDAF'S
* Worked on Oozie workflow engine for job scheduling.

**UnitedHealth Group, Basking Ridge, NJ**

**Hadoop Developer,** Jan 2012 to July 2013

**Description**:

UnitedHealth Group is the most diversified health care company in the United States and a leader worldwide in helping people live healthier lives and helping to make the health system work better for everyone.

The particular project involves in assigning a personal physician for each and every customer respective to particular area. Categorizing the diseases encountered frequently based on the claims and on the geographical location. The particular project also analyses the effectiveness of campaigns conducted at various places.

**Responsibilities:**

* Developed multiple **Map Reduce** jobs in Java for data cleaning and preprocessing
* Responsible for design and creation of **Hive** tables, partitioning, bucketing, loading data and writing hive queries.
* Created **HBase** tables to store various data formats of personally identifiable information (PII) data coming from different portfolios.
* Wrote **Pig** scripts to transform the data in **HDFS**
* Responsible to manage data coming from different sources.
* Involved in managing and reviewing **Hadoop log files**.
* Installed **Oozie** workflow engine to run multiple **Hive** and **Pig** jobs.

**Environment:**

Apache Hadoop 2.x, MapReduce, HDFS, Hive, HBase, Pig, Oozie, Sqoop, Linux, Oracle 11g, Java 7, Eclipse

**Metlife, Bridgewater, NJ**

**Sr. Java Developer,** Oct 2010 – Dec 2011

**Description:** Help Clinical Desktop (CD) is a web application based on J2EE architecture.  It provides web-based access to inpatient and outpatient data from the IHC Clinical Data Repository (CDR). Users can review as well as update or enter new data through CD. Many clinicians use CD daily for accessing patient data, which aids in making important clinical decisions. CD is mainly composed of a “shell” or “core”, and various modules running inside of the “shell”. Intended audiences for this document are mainly programmers.

**Responsibilities:**

* Full life cycle experience including requirements analysis, high level design, detailed design, UMLs, data model design, coding, testing and creation of functional and technical design documentation.
* Used Spring Framework for MVC architecture with Hibernate to implement DAO code and also used Web Services to interact other modules and integration testing.
* Developed and implemented GUI functionality using JSP, JSTL, Tiles and AJAX.
* Designed database and involved in developing SQL Scripts.
* Used SQL navigator as a and involved in testing the application.
* Implementing the Design Patterns like MVC-2, Front Controller, Composite view and all Struts framework design patterns to improve the performance.
* Used Clear case, and also subversion for maintaining the source version control.
* Wrote Ant scripts to automate the builds and installation of modules.
* Involved in writing Test plans and conducted Unit Tests using JUnit.
* Used Log4j for logging statements during development.
* Design and implementation of log data indexing and search module, and optimization for performance and accuracy. To provide a full text search capability for archived log data, utilizing Apache Lucene library.
* Involved in the testing and integrating of the program at the module level.
* Worked with production support team in debugging and fixing various production issues.

**Environments:**

Java1.5,AJAX,XML,Spring3.0,Hibernate2.0,Struts1.2,Webservices,Websphere7.0,Junit,Oracle10g,SQL, PL/SQL, log4j, RAD 7.0/7.5, Clear case, Unix, HTML, CSS, Java script.

**Client: Ramco Systems, Chennai, INDIA**

**Java Developer** June 2009 -September 2010

Ramco Systems is a fast-growing enterprise software player disrupting the market with its multi-tenanted cloud and mobile-based enterprise software in the area of HCM and Global Payroll, ERP and M&E MRO for Aviation

**Responsibilities:**

* Coordinated with the business analysts, project managers to analyze new propose Ideas/Requirements, designed the integrated tool, developed and implemented all the modules.
* Designed database and involved in developing SQL Scripts.
* Used Case Studio for developing the DB Design and generating SQL file for various databases.
* Contributed significantly in designing the **Object Model** for the project as senior developer and Architect.
* Responsible for development of Business Services.
* Developed Business Rules for the project using Java.
* Developed portal screens using **JSP, Servlets,** and **Struts** framework.
* Developed the test plans and involved in testing the application.
* Implementing the Design Patterns like MVC-2, Front Controller, Composite view and all Struts framework design patterns to improve the performance.
* Re-engineered OMT Wholesale Internet Service Engine (WISE) using an “n” tiered architecture involving latest technologies like **EJB**, **XML** and **JAVA**.
* Used CVS for maintaining the source version control.
* Used **Log4j** for logging statements during development.
* Design and implementation of log data indexing and search module, and optimization for performance and accuracy. To provide a full text search capability for archived log data, utilizing Apache Lucene library.
* Used **Spring Framework** for MVC architecture with **Hibernate** to implement DAO code and also used **Web Services** to interact other modules and integration testing.
* Developed and implemented GUI functionality using **JSP, JSTL, Tiles and AJAX.**
* Designed database and involved in developing SQL Scripts.
* Used SQL navigator as a tool to interact with DB Oracle 10g.
* Developed portal screens using **JSP, Servlets,** and **Struts** framework
* Involved in writing Test plans and conducted Unit Tests using **JUnit**.
* Used **Log4j** for logging statements during development

**Environments**:

Java, J2EE, Struts 1.2/2.0, JDK, JSP, Servlets, EJB 3.0, Java Beans, JavaScript, HTML XML, Eclipse, CORBA SSL, JUnit, Log4j, CVS, Deployment in Web Logic.