**Aman Pathak – Big Data/Python**

**Contact Details**

Phone: 4433963910

Email: [aman@techmail2.com](mailto:aman@techmail2.com)

LinkedIn: www.linkedin.com/in/aman-patthak

**Summary**

* Seasoned Big Data professional with over 10 years of experience in managing and processing high-volume data sets using Python, Apache Hive, Spark, PySpark, Apache Hadoop, SQL, Machine Learning, and Shell Scripting.
* Expertise in Big Data technologies and Hadoop ecosystem components such as Spark, HDFS, MapReduce, Pig, Hive, YARN, Sqoop, Flume, Kafka, and NoSQL systems like HBase, Cassandra.
* Proficient in Kafka streaming, administration, and monitoring, along with writing MapReduce Jobs in Java for processing structured and semi-structured data.
* Proficient in developing data transformation and analytical applications in Spark, Spark-SQL using the Scala programming language, and creating real-time data streaming solutions using Apache Spark/Spark Streaming, Kafka.
* Experienced in managing Databricks workspace for Business analytics, managing clusters in Databricks, and managing the machine learning life cycle with MLlib.
* Hands-on experience in developing Custom UDFs using Java for use in Pig and Hive, and designing and querying NoSQL databases like HBase, Cassandra.
* Worked with Java API, Rest API for real-time analytics on HBase data, and streaming data using Apache Kafka, Storm, and Flume.
* Proficient in Java/J2EE Technologies like Servlets, JSPs, Hibernate, SpringDI, Struts, and JPA with knowledge of SDLC concepts, and configuring RedShift, Elasticsearch, and DynamoDB with EC2 Instances.
* Experienced in working with Machine learning libraries (spark-MLlib) and implementing ML algorithms for clustering, regression filtering, and dimensional reduction.
* Involved in requirement analysis, reviews, working sessions to understand requirements & system design, and experienced in using Agile methodologies including extreme programming, Scrum Process, and Test-Driven Development (TDD).
* Proficient in creating data ingestion pipelines, data transformations, data management, and real-time streaming at an enterprise level, and working with AWS technologies like S3 and EMR for storage, Big data processing, and analysis.
* Have good experience on Web, Client/Server technologies using Java, J2EE, Servlets, JSP, EJB, JDBC, and BI Tools like Tableau and Plotly.
* Proactive and well-organized with effective time management skills and problem-solving skills.
* Strong understanding of SQL and database concepts, with experience in working with various databases and data formats.
* Experienced in using version control tools like Bit-Bucket, GIT, SVN, and CVS, and build scripts using Maven, ANT, and Gradle.
* Proficient in working with various file formats like delimited text files, click stream log files, Apache log files, Avro files, JSON files, XML Files, and understanding of compression techniques used in Hadoop processing like Gzip, SNAPPY, LZO.
* Experienced in working with cloud environments like Amazon Web Services (AWS) EMR, EC2, ES, and S3, and Microsoft Azure.

**Technical Skills**

|  |  |
| --- | --- |
| **Big data Ecosystem** | Hadoop, MapReduce, Apache Hive, Apache Spark (PySpark), Apache Kafka, Apache HBase, Apache Cassandra, Apache Nifi, Apache Kudu, Apache Oozie, Apache Zookeeper, Sqoop, Cloudera, Elasticsearch, Log4j |
| **Programming Languages** | Python, C++, SQL, Shell Scripting, Scala, Java, J2EE |
| **Cloud Technologies** | AWS (EC2, S3, EMR, Redshift, Athena, Kinesis, Glue, Lambda, Dynamo DB, Quick sight), Azure, Terraform |
| **SQL Databases** | Oracle, MySQL, DB2, Teradata, PostgreSQL |
| **NoSQL Databases** | HBase, Cassandra, MongoDB |
| **BI Tools** | Tableau, Power BI, Qlik Sense |
| **Project Management Tools** | JIRA, Rally |
| **IDEs and Editors** | IntelliJ, Eclipse, Jupyter, PyCharm, VS Code, Sublime Text, Atom, Google Collab) |
| **Version Control** | GitHub |
| **Operating System** | UNIX, Linux, Windows, MAC OS |
| **Methodologies** | Agile, Scrum |
| **Machine Learning** | MLlib, TensorFlow, Keras, PyTorch, Pandas, Scikit-Learn |

**Education**

* Bachelor of Science in Computer Science and Bachelor of Management with emphasis on International Business | Webster University
* Honors | GPA: 3.8

**Professional Experience**

Anthem, Phoenix, AZ **Feb 2022 – July 2023**

**Big Data Developer**

**Responsibilities:**

* Implemented Apache Spark with PySpark and Spark SQL for efficient data testing and processing, improving data processing speed.
* Used SQL for database querying, validation, and mapping activities, which facilitated the extraction of meaningful insights from healthcare data.
* Employed Spark transformations and actions using PySpark for data cleansing, improving data quality and consistency.
* Developed linear regression models for continuous measurements on healthcare data using the Spark with PySpark API, aiding in predictive analysis.
* Used Spark and Spark-SQL to read parquet data and create tables in Hive using the PySpark API.
* Performed sorting, joining, aggregating, filtering, and other transformations on datasets using Spark, enabling more accurate data analysis.
* Converted Hive/SQL queries into Spark transformations using Spark RDDs, enhancing data processing speed.
* Managed data from different sources using Spark, PySpark, and Spark SQL, ensuring seamless data integration.
* Created internal and external Hive tables, utilizing static and dynamic partitions, and bucketing for improved efficiency, which optimized query performance.
* Developed shell scripts to automate the creation of Hive tables and data loading, reducing manual effort and increasing efficiency.
* Analyzed business requirements in the healthcare industry and collaborated with the team to align data solutions with set goals.
* Assisted in maintaining database systems to meet organizational needs and ensure data integrity.
* Conducted extensive data validation using complex SQL queries, performed back-end testing, and addressed data quality issues, significantly enhancing the accuracy of the healthcare data.
* Integrated Maven build into workflows to automate the build and deployment processes, increasing deployment efficiency.
* Processed various formats of flat files and stored them as different partition models in HDFS, enhancing data storage efficiency.
* Developed PySpark scripts, tested their performance, and shared the optimized scripts across modules, improving overall team productivity.
* Monitored Spark jobs and logged activities using Spark application master, ensuring smooth data processing workflows.

Environment: Hadoop, Apache Spark, Spark-SQL, AWS, Maven Build, Rally, IntelliJ, HDFS, Map Reduce, Hive, Kafka, Sqoop, PySpark, Talend, Cloudera, Linux, SQL, GitHub, Jenkins, Terraform, UNIX, NIFI, HBase, AWS EC2, YARN, RDBMS, Zookeeper.

**Ford Direct MI Mar 2021 – Jan 2022**

**Big Data Developer**

**Responsibilities:**

* Developed and executed PySpark projects, enhancing data testing and processing capabilities.
* Constructed a linear regression model using PySpark to optimize wind turbine data observation, leading to more accurate data insights.
* Worked extensively with the Spark ecosystem, utilizing PySpark and Spark SQL on various data formats like text files and CSV files, improving data processing efficiency.
* Managed Hadoop clusters using the Cloudera Manager tool, ensuring optimal performance and stability.
* Developed user interfaces using JSP, JavaScript, and HTML with the Backbone.js framework, improving user experience.
* Transferred data from source systems to Azure data systems using Data Factory and Spark SQL, enhancing data accessibility and integration.
* Deployed web and enterprise Java components, improving system functionality and user interaction.
* Utilized Hibernate ORM tools for automating the mapping between SQL databases and Java objects, reducing manual effort.
* Used Log4j logging API for effective debugging and monitoring of Java applications.
* Implemented Spark API over Hortonworks Hadoop YARN for comprehensive data analytics in Hive.
* Utilized Azure Data Lake Analytics and Data ingestion to various Azure services, processing data in Azure Databricks, improving data processing speed.
* Imported data from AWS S3 to Spark RDD, performing transformations and actions on RDDs, enhancing data processing efficiency.
* Utilized cloud provisioning tools such as Terraform and CloudFormation, improving infrastructure management efficiency.
* Implemented monitoring solutions in Ansible, Terraform, Docker, and Jenkins, reducing system downtime.
* Implemented Lambda architecture using Azure Data platform like Azure Data Lake, Data Factory, enhancing data processing and storage capabilities.
* Created and maintained highly scalable, fault-tolerant multi-tier AWS and Azure environments across multiple availability zones using Terraform.
* Provisioned Kubernetes clusters on AWS for creating development and testing environments of different applications using Docker, Ansible, and Terraform, reducing setup time.

Environment: Hadoop, Map Reduce, Hive, PySpark, SQL, DB2, GitHub, Jenkins, Terraform, UNIX, Cloudera, Kafka, Sqoop, Java, Scala, NIFI, HBase, AWS EC2, S3, HDFS, YARN, RDBMS, Zookeeper, NoSQL, Cassandra, Python.

**Northern Natural Gas, Omaha, NE Feb 2020 – Feb 2021**

**Hadoop/Spark Developer**

**Responsibilities:**

* Leveraged PySpark for accelerated data testing and processing, managing data from various sources, and performing data quality checks.
* Utilized Kafka for real-time data fetching and processing, with business logic implemented using Kafka Direct Stream in Spark Streaming.
* Imported data from relational databases like MySQL, Oracle using Sqoop, and handled the import of both structured and unstructured data into HDFS.
* Developed MapReduce jobs for data cleaning and validation and implemented Hive Partitioning and Bucketing on the collected data in HDFS.
* Set up Sqoop jobs for large data exchanges between RDBMS and Hive clusters and used Zookeeper as a backup server and job scheduler for Spark Jobs.
* Worked with Cloudera distribution, deploying on AWS EC2 instances, and connected Cassandra database to Amazon EMR File System for storing the database in S3.
* Implemented Amazon EMR for processing Big Data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Configured workflows involving Hadoop actions using Oozie and collaborated with the SCRUM team in delivering agreed user stories on time for every sprint.
* Utilized LINUX and UNIX operating systems, demonstrating good knowledge of Linux Shell scripts.
* Wrote Terraform scripts for application alerts tracking and built and implemented real-time streaming ETL pipeline using Kafka Streams API.
* Worked with Hive to implement web interfacing, storing data in Hive tables, and migrated Map Reduce programs into Spark transformations using Spark and Scala.

**Environment**: Hadoop, Map Reduce, Hive, Spark, Oracle, Terraform, GitHub, Tableau, UNIX, Cloudera, Kafka, Sqoop, Scala, NIFI, HBase, AWS EC2, S3, MySQL, LINUX, HDFS, ETL, YARN, RDBMS, Zookeeper, NoSQL, Cassandra, Oozie, Python.

**SEIMENS, MD Jan 2019 - Jan 2020**

**Sr Hadoop / Spark Developer**

**Responsibilities:**

* Assisted in data migration from various conventional data sources to Hadoop ecosystem using Apache Kudu and Spark.
* Managed data from disparate sources, loading and transforming large sets of structured, semi-structured, and unstructured data.
* Utilized Hive Script in Spark for data cleaning and transformation purposes.
* Developed applications to map data between different sources and destinations using Python and Scala.
* Loaded data into Hive partitioned tables and exported analyzed data to relational databases using Kudu for visualization and generating reports for the BI team.
* Used Apache Nifi for file conversions and data processing.
* Collaborated with data modelers and other developers during implementation.
* Worked with AWS infrastructure services like Amazon Simple Storage Service (Amazon S3) and Amazon Elastic Compute Cloud (Amazon EC2).
* Utilized ELK stack and Kafka services for data processing and analysis.
* Explored new Elasticsearch components to improve the existing system, such as Elastic Cloud Enterprise for centralizing all clusters, playbooks, Ansible, and machine learning.
* Installed, configured, administered, and supported multiple Kafka and Elasticsearch clusters; performed maintenance and troubleshooting, capacity planning, and growth projections.

**Environment**: CentOS 7, Python, Sqoop, Hive, Scala, AWS, CDH 5.12.1, Kudu, Spark, Oozie, Elastic, Logstash, Kibana, Kafka Streaming, Confluent, Cloudera Manager, Hue, SQL Server, Microsoft Azure DevOps, Git, Agile Methodology, ETL, NIFI, Amazon EC2, S3, HDFS, Tableau, JIRA.

**Standard Charted Bank, NJ May 2016 - Oct 2018**

**Hadoop Developer**

**Responsibilities**:

* Developed and implemented ETL processes for data transformation and cleaning using Hive and Python, enhancing data quality and efficiency.
* Utilized Kafka for real-time data ingestion into Spark Streaming, processing the data in Parquet format for improved data accessibility.
* Worked with HiveQL for data analysis, implementing business logic into Hive Queries, and creating Hive tables for structured data storage.
* Used Python for automating data movements and for pattern matching in build logs, streamlining data processing workflows.
* Imported and exported data into HDFS from Oracle Database and vice versa using Sqoop, ensuring seamless data integration.
* Developed MapReduce jobs for data cleaning, validation, and transformation, improving data reliability.
* Managed data migration from RDBMS to HDFS using Sqoop, facilitating large-scale data transfers.
* Used Amazon Redshift for data warehousing and Tableau for data visualization, providing actionable insights for business decision-making.
* Collaborated with BI teams to design ETL workflows and generate reports, supporting data-driven decision-making.
* Worked on partitioning Hive tables and running scripts in parallel to reduce the runtime, optimizing data processing speed.
* Handled bug fixes during the QA phase, supported QA testing, and identified the sources of defects, ensuring the delivery of high-quality solutions.
* Prepared detailed specifications for ETL processes and functional documents for data transformation workflows, providing clear guidelines for project execution.
* Participated in Agile methodologies, daily Scrum meetings, and Sprint planning, promoting efficient project management.

Environment: Hadoop, MapReduce, HDFS, AWS, Hive, Java, SQL, Agile, Cloudera Manager, Spark, Pig, Python, PostgreSQL, Sqoop, EMR, Oozie, HBase, Zookeeper, PL/SQL, MySQL, DB2, Teradata, Kafka Administration, Kafka Monitoring, Amazon Redshift, Oracle, NoSQL, HBase, MongoDB, Scala, Impala.

**Credit One Bank , Nevada Aug 2014 – Dec 2015  
Hadoop Developer**

**Responsibilities**:

* Developed MapReduce programs in Java to process large datasets, improving data processing efficiency.
* Implemented Kafka Direct Stream in Spark Streaming to manage real-time data ingestion from weblogs, enhancing the timeliness and accuracy of data analysis.
* Used HBase to handle time-series data, enabling efficient storage and retrieval of sequential data.
* Integrated HBase with MapReduce to move bulk amounts of data into HBase, streamlining data storage processes.
* Imported data from Oracle to HDFS using Sqoop, ensuring data availability for Hadoop processing.
* Developed Pig scripts for data cleaning and processing, contributing to data quality and readiness for analysis.
* Installed and configured Hive, and wrote Hive UDFs to implement business requirements, enhancing data querying capabilities.
* Optimized Hive queries by partitioning Hive tables and running scripts in parallel, reducing runtime.
* Developed Scripts and Batch Jobs to schedule various Hadoop Programs using Oozie, automating data processing tasks.
* Converted data from Avro to Text file format using Pig scripts, facilitating data readability and analysis.
* Used Avro Storage within Pig Latin to load and store data, supporting diverse data formats.
* Participated in the development of low-level documents, focusing on functional and non-functional requirements, to guide the creation of data processing systems.
* Prepared functional specification and deployment instruction documents, providing clear guidelines for system implementation and use.
* Collaborated with QA teams to troubleshoot and fix defects, ensuring smooth and error-free code deployment.

Environment: Hadoop, Kafka, Sqoop, MapReduce, HDFS, Hive, Pig, HBase, Linux, XML, Java, Eclipse, Oracle, JIRA, GitHub, CDH4, Rest API, Spark Streaming, Oozie.

**BCBS, NJ Dec 2013 - Jul 2014**

**Jr. Data Analyst**

**Responsibilities**:

* Participated in all phases of data mining, including data collection, data cleaning, model development, validation, and visualization.
* Developed Tableau dashboards of Key Performance Indicators (KPIs) to provide real-time aggregated data views to decision-makers.
* Assisted in the development of reports, data models, and dashboards by gathering and transforming data into meaningful insights to drive business recommendations and decision-making using Tableau.
* Developed interactive data visualizations in Tableau using both relational and aggregate data sources.
* Created complex dashboards in Tableau using parameters, sets, groups, and calculations, and implemented drill-down and drill-up functionalities in worksheets, further customizing with filters and actions.
* Involved in presenting data using a variety of visualization types such as crosstab scatter plots, geographical maps, pie charts, bar charts, motion charts, small multiples, heatmaps, and density charts.
* Gained experience in using SSIS to create ETL packages for validating, extracting, transforming, and loading data into data warehouse databases.
* Designed and implemented multiple dashboards using a customized business intelligence tool for in-house metrics.

Environment: SQL Server, Oracle, SAS, Tableau, SSIS, ETL.