|  |
| --- |
| ABHISHEK PATHAK |
| Vancouver, BC | 437-371-8781 | abhishekabhi000@gmail.com | [Linkedin.com/in/](http://linkedin.com)abhishek-pathak-3bb429b | |

|  |
| --- |
| **SUMMARY** |
| * 10+ years of experience in SDLC with key emphasis on the trending Big Data Technologies - Spark, Scala, Java, Hadoop, Tableau, Cassandra, Kafka, Mongo DB, Airflow, PagerDuty, RedShift * Good Knowledge of Big Data and Data Warehouse Architecture and Designing Star Schema, Snowflake Schema ,Fact and Dimensional Tables, Physical and Logical Data Modeling. * Architect, design & develop Big Data Solutions practice including set up Big Data roadmap, build supporting infrastructure and team to provide Big Data. * Architecting and implementing Portfolio Recommendation Analytics Engine using Hadoop MR, Oozie, Spark SQL,Spark Mlib and Cassandra. * Excellent understanding of Hadoop architecture and underlying framework including storage management. * Extensive experience in data modeling, data architect, solution architect, data warehousing & master data management (MDM) concepts. * Expertise in architecting Big data solutions using Data ingestion, Data Storage. * Experienced in Worked on NoSQL databases - HBase, Cassandra & MongoDB, database performance tuning & data modeling. * Architecting, Solutioning and Modeling DI (Data Integrity) Platforms using Sqoop, flume, Kafka, Spark Streaming, Cassandra. * Excellent understanding of Hadoop architecture and underlying framework including storage management. * Extensive experience in data modeling, data architect, solution architect, data warehousing & master data management (MDM) concepts. * Expertise in architecting Big data solutions using Data ingestion, Data Storage. * Experienced in Worked on NoSQL databases - HBase, Cassandra & MongoDB, database performance tuning & data modeling. * Architecting, Solutioning and Modeling DI (Data Integrity) Platforms using Sqoop, flume, Kafka, Spark Streaming, Cassandra. * Design, implement and maintain all AWS infrastructure and services within a managed service environment. * Design, Deploy and maintain enterprise class security, network and systems management applications within an AWS environment. * Implement process and quality improvements through task automation. Institute infrastructure as code, security automation and automation or routine maintenance tasks * Perform data migration from on premises environments into AWS. * Support the business development lifecycle (Business Development, Capture, Solution Architect, Pricing and Proposal Development) * Strong knowledge of Amazon Kinesis, AWS Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS), and Amazon Simple Workflow Service (Amazon SWF) * Strong knowledge with Web Services, API Gateways and application integration development and design   **Technical Skills:**   |  |  | | --- | --- | | Hadoop/Big Data Technologies | HDFS, Sqoop, Flume, Pig, Hive, Oozie, Spark, Zookeeper, Kafka. | | AWS | S3, Athena, Glue | | Hadoop Distribution | knowledge of Horton Works, AWS | | Python Libraries | Pandas, Numpy, Matplotlib, Scikit-learn, pyathena, boto3 | | Data Visualization | Tableau, pyplot, seaborn, ggplot, Excel | | Programming & Scripting | Python, Pyspark, Java, R, Pyspark and SQL (Presto, SQL Server), shell-script | | Machine Learning Algorithms | Linear and Logistic Regression, Clustering, SVM, Random Forest, Ensemble Methods, Neural Networks | | Databases | Oracle, MY SQL, MS SQL server, PL/SQL Developer, HBase | | Version Control | SVN, GIT | | Operating Systems | Linux, Unix, Mac OS-X, Cen OS, Windows10, Windows 8, Windows 7 | | Other Tools: | PyCharm, Jupyter Notebook, Eclipse, Microsoft Word, PowerPoint, JIRA, ServiceNow. |      |  | | --- | | **APRIL 2019 – CURRENT**  **SDE, ICBC Vancouver**   * Leading the migration of data pipelines from to Hadoop ecosystem from existing Oracle world. * Developing config driven framework to utilize common code base across different pipelines. * Migration of Scheduling from Crontab to CTRL.M * Designing system which will allow onboarding of third-party data into internal eco system. * Designing Data Quality Framework to support ad-hoc as well as scheduled execution of Athena-Sql queries with Airflow. * Developing data quality Dashboard on AWS QuickSight and integrating with DQ framework. * Redesigning existing pipelines to reduce the ingestion time from 40 minutes to under ~ 15 minutes. * Leading the Tech Stack upgrade on AWS from EMR 600 based spark/scala/Hudi to EMR 660. * Playing central role in setting the team culture of code reviews/ design reviews. And being the gate keeper to make sure across the team , best distributed design patterns are followed. * Enhancing the existing CI CD pipelines from EC2 to ECS based distributed pattern to reduce the downtime by 99% and reduce the build time by 50% * Partnered with ETL developers to ensure that data is well cleaned, and the data warehouse is up to date. * Selected and generated data into CSV files and stored them into AWS S3 by using AWS EC2 and then structured and stored in AWS Redshift. * Used AWS S3 Buckets to store the file and injected the files into Snowflake tables using Snow Pipe and run deltas using Data pipelines. * ETL Development: Demonstrated ability to design and implement complex ETL workflows using Informatica PowerCenter, ensuring efficient data extraction, transformation, and loading from various data sources into target systems. * Migrate data from on-premises to AWS storage buckets. * Developed a python script to transfer data from on-premises to AWS S3. Developed script that will hit REST API's and extract data to AWS S3. * Deployed application using AWS EC2 standard deployment techniques and worked on AWS infrastructure and automation. Worked on CI/CD environment on deploying application on Docker containers. * Worked on Ingesting data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue and Step Functions. * Performance Tuning and Optimization: Implemented performance optimization techniques in Informatica PowerCenter, including parallel processing, partitioning, and indexing, to enhance data processing speed and improve overall ETL performance. * Extensively used ETL for supporting data extraction, transformations and loading processing, in a complex EDW using Talend/Data stage. * Designing and implementing data storage solutions on Azure. * Extensive experience in designing, developing, and deploying ETL (Extract, Transform, Load) solutions using SSIS for data integration and data warehousing projects. * Proficient in creating SSIS packages to extract data from various data sources such as databases, flat files, Excel spreadsheets, and RESTful APIs. * Developed complex data transformation logic using SSIS components including Data Flow Task, Control Flow Task, Script Task, and Execute SQL Task. * Implemented error handling and data quality checks within SSIS packages to ensure data integrity and accuracy.   **JUNE 2018 – APRIL 2019**  **SDE, CerebriAI**   * Enhanced and optimized Spark/Scala/ pyspark jobs to aggregate, group and run data mining tasks using the Spark framework. * Worked on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and controlling and granting database access and Migrating on premise databases to Azure Data lake store using Azure Data factory * Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns. * Informatica PowerCenter Expertise: Possess extensive experience and expertise in Informatica PowerCenter, showcasing proficiency in designing, developing, and deploying ETL (Extract, Transform, Load) processes using the tool. * Implemented data cleansing, data validation, and data enrichment tasks using SSIS transformations and external data sources. * Implemented scheduling and automation of SSIS packages using SQL Server Agent or third-party scheduling tools. * Conducted unit testing and system testing of SSIS packages to ensure proper functioning and adherence to business requirements. * Worked with database administrators to deploy SSIS packages to production environments and troubleshoot deployment issues. * Error Handling and Exception Management: Developed robust error handling mechanisms in Informatica PowerCenter to capture and manage data transformation errors, ensuring data integrity and enabling effective error resolution. * Advanced skills in Microsoft SQL, coupled with practical experience using Entity Framework for efficient data handling. * Built Teradata ELT frameworks which ingests data from different sources using Teradata Legacy load utilities. * Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards. * Importing and exporting data into HDFS and hive using Sqoop and Kafka with batch and streaming. * Design and developed services to persist and read data from Hadoop, HDFS, Hive and writing Java based MapReduce batch jobs using Hadoop Data Platform. * Involved in complete Big Data flow of the application data ingestion from upstream to HDFS, processing the data in HDFS and analyzing the data using several tools. * Cloud Integration: Extensive experience in integrating Informatica PowerCenter with cloud-based platforms such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP), Informatica leveraging services like AWS Glue, Azure Data Factory, or GCP Dataflow. * Imported the data from various formats like JSON, ORC and Parquet to HDFS cluster with compressed for optimization. * Experienced on ingesting data from RDBMS sources like - Oracle, SQL Server and Teradata into HDFS using Sqoop. * Deployed Pyspark applications and developed in Databricks cluster. * Experience in managing and reviewing huge Hadoop log files. * Developed exposure with multiple BI and analytic tools, including but not limited to Tableau, Qlik, Cognos and SAS with an emphasis on the creation of visualizations, dashboards, and interactive reports. * Worked with Client support environment assisting to troubleshoot SAS queries and collaborate with SAS Admin to diagnose user identified issues.   **NOVEMBER 2017 – JUNE 2018**  **SDE, Groupon**   * Leading the redesigning of data pipeline, leading to a significant reduction in execution time with lesser resources. Assisting in the ongoing feature development of our Data quality framework. * Designed and developed the end to end click stream pipeline with spark, kafka & Mysql. Playing a pivotal role in shaping team culture and promoting social cohesion. * Collaborate with others in crafting and implementing technical vision.Provide technical leadership to a strong team of data scientists, engineers, and analysts.Build cloud agnostic, highly scalable, big data, machine-learning applications on Kubernetes. * Re-designed and developed a critical ingestion pipeline to process over 300 TB of data. * Tech Stack: Spark, Scala, Cassandra, Python, MonetDB, Airflow. Docker, K8, Azure. * Responsible for requirements gathering and analyzing the data sources. * Responsible for building scalable distributed data solutions using Hadoop. * Worked very closely with the Hadoop Administrator to set up the Hadoop cluster. * Worked on analyzing Hadoop cluster and different Big Data analytic tools including Pig, Hive HBase database and SQOOP. * Importing and exporting data into HDFS and Hive using SQOOP. * Installed Hadoop, Map Reduce, HDFS, and Developed multiple map reduce jobs in PIG and Hive for data cleaning and pre-processing. * Real-time Data Integration: Implemented Informatica PowerCenter Real-Time Edition to enable real-time data integration and streaming, ensuring timely and accurate data delivery for operational reporting and analytics. * Generate OBIEE reports to verify the Hive table's data. * Used Pig as ETL tool to do Transformations, even joins and some pre-aggregations before storing the data onto HDFS. * Experience working in spark eco system using Spark-Sql and Scala queries on different formats like Text file, CSV file. * Involved in loading data from LINUX file system to HDFS. * Load and Transform large sets of structured and semi structured data. * Exported data from HDFS environment into RDBMS using Sqoop for report generation and visualization purpose. * Worked on creating dependencies of activities in Azure Data factory * Created Hive tables to store variable data formats of PII data coming from different portfolios. * Implemented a script to transmit sysprin information from Oracle to Hive using Sqoop. * Designed and developed MapReduce jobs to analyze the data. * Implemented best income logic using Pig scripts and UDFs. * Worked with application teams to install operating system, Hadoop updates, patches, version upgrades as required. * Responsible to manage data coming from different sources. * Involved in loading and transforming large sets of structured, semi structured, and unstructured data. * Installed Oozie workflow engine to run multiple Hive and pig jobs. * Worked in Base SAS programming including Data steps, PROC SQL, Advanced prompting. * Supported in setting up QA environment and updating configurations for implementing scripts with Pig and Sqoop. * Involved in email campaign launch and served as point of contact for anything related to data, pulling customers list, matching and finished training in Strong View email platform software.     **JUNE 2016 – NOVEMBER 2017**  **Co-Founder, Woot**   * Beta had 10000 users and 50000 videos. Worked with AWS SQS, Flume, S3, NodeJs, Java, Spark, Redis,Cassandra, Mongo, AWS RDS, Facebook Add Network. Click Stream data pipeline development with Spark, SQS, * Flume & Cassandra and backend API stack in NodeJs. Coding, product roadmap, system administration,development, testing, emails, pitching, team planning * Development and design of backend for booking portal. Set up the backend consisting of Python + Django,RabbitMQ, Postgres, Docker, Jenkins, Nginix & Gunicorn. Designed and developed invoicing platform after advent of GSTIN. * Developed Spark/Scala, Python for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources. * Data sources are extracted, transformed, and loaded to generate CSV data files with Python programming and SQL queries. * Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself. * Experience in modeling and quantification concepts including statistics and probability principles * Experience with Azure Analysis Service data modeling in Visual Studio/SSDT (SQL Server Data Tools) * Designed, developed and did maintenance of data integration programs in a Hadoop and RDBMS environment with both traditional and non-traditional source systems as well as RDBMS and NoSQL data stores for data access and analysis.   **SEPTEMBER 2015 – MAY 2016**  **Senior Software Engineer, SAP Lab**   * Developed Kafka as a Service on OpenStack End to End design and development of on demand Kafka as service including Alerting & Monitoring with ELK and Grafana. Designed Log monitoring pipeline with ELK stack and JMX, Riemann, InfluxDb & Grafana. * Paper submission to CloudFoundry Summit : proposal for cloud foundry functions for server less compute and a working prototype based on Rediss, NodeJs, CloudFoundry. Setup continuous integration & deployment tool. Grooming freshers into CloudFoundry landscape ,delivered multiple training sessions. | |  |   **JUNE 2013 – SEPTEMBER 2015**  **Senior Software Engineer, British Telecom**   * Proposed and migrated existing ETL pipelines from Ab initio to Hadoop ecosystem. Reduced the licensing cost by more than 50%. Designed and developed data pipelines with Sqoop, Flume, Pig, Java, Mapreduce, Hive, HDFS.   **DECEMBER 2012 – JUNE 2013**  **Senior Software Engineer, RIGHTSTER**   * Implemented an on demand, config driven data pipeline over AWS EMR. Workflow engine to manage horizontal and vertical dependencies between different pipelines. Coded in Java, Python, Pig, EMR, Json.   **FEB 2010 – NOVEMBER 2012**  **Lead Software Engineer, SAMSUNG Electronics**   * Implemented product features for Android Enterprise Suit eg. Email, Calendar, Contacts(Active Sync features).Incepted and developed end to end, SharePoint application for Android devices, launched in Samsung Market place.     **SEPTEMBER 2008 – FEB 2010**  **Software Engineer, SASKEN Communications**   * Developed testing framework and BSC simulator saving BSC hours per day. * Tech Stack: C, C++, gcc.   **JUNE 2006 – MAY 2008**  **Intern, NIT Rourkela**   * Intrusion Detection and Prevention: Build rule-based Firewall to filter bad incoming traffic to university networks.   **JUNE 2006 – MAY 2008**  **Summer Intern, ELITECORE Technologies**   * Analysing firewall traffic to generate digital signatures for intrusion events.  |  | | --- | | **EDUCATION** | | **2008**  **National Institute of Technology - India**  **Bachelor Of Technology, Electronics** | |