GLENN KHAN

(929) 435 6965

Gbigdata1983@gmail.com

US Citizen, New York City, NY

PROFESSIONAL SUMMARY

- Over **8+** years of extensive IT experience and more than **6** years of experience as a Hadoop development, operations and management of small to medium clusters using distributions like CDH, HDP and ECS.
- Hands on Experience in Installing, configuring and using Hadoop Ecosystem components like HDFS, Hadoop MapReduce, Yarn, Zookeeper, Sqoop, Impala, Flume, Hive, Pig, HBase, Spark, Pig, Oozie.
- Involved in converting HQL queries into Spark transformations using Spark RDDs.
- Experienced in installing and configuring Flume, Hive, Pig, Sqoop and Oozie on the Hadoop cluster.
- Experience in deploying a Hadoop clusters using Cloudera 5.X integrated with Ambari for management, monitoring and alerting.
- Experienced in launching and setting up of Hadoop clusters on AWS as well as physical servers, which includes configuring different Hadoop components
- Developed and monitored of Puppet Configuration Manager and automated configuration of Hadoop ecosystem.
- Good knowledge on implementation and design of big data pipelines and implementing of ETL/ELT processes
- Hands on experience in using MapReduce programming model for Batch processing of data stored in HDFS.
- Responsible for designing and building a Data Lake using Hadoop and its ecosystem components.
- Very good experience in complete project life cycle (design, development, testing and implementation) of client-server and web applications
- Developed Spark Applications by using Scala, Java and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources
- Worked with the Spark for improving performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD's, Spark YARN
- Experienced in Apache Spark for implementing advanced procedures like text analytics and processing using the in-memory computing capabilities written in Scala
- Experience using middleware architecture using Sun Java technologies like J2EE, Servlets, and application servers like Web Sphere and Web logic.
- Used Different Spark Modules like Spark core, Spark RDD's, Spark Data frame, Spark SQL.
- Converted Various Hive queries into Spark transformations and actions that are required.

- Experienced in working on apache Hadoop open source distribution with technologies like HDFS, Map-reduce, Python, Pig, Hive, Hue, HBase, SQOOP, Oozie, Zookeeper, Spark, Spark-Streaming, Storm, Kafka, Cassandra, Impala, Snappy, Green Plum and MongoDB, Mesos.
- Experienced in Tableau and Spotfire and enabled the JDBC and ODBC connectivity from those to Hive tables.
- Designed neat and insightful dashboards in Tableau.
- Experienced in installing and configuring Kerberos for the authentication of users and Hadoop daemons.
- Hands on experience in Linux Hadoop activities on RHEL &Cent OS.
- Knowledge on Cloud technologies like AWS Cloud.
- Experienced in benchmarking, backup and disaster recovery of NameNode Metadata.
- Experienced in working with popular frame works like Spring MVC and Hibernate.
- proficient in source code management tools Git.
- Excellent interpersonal and communication skills, creative, research-minded with problem solving skills.
- Ensure that critical customer issues are addressed quickly and effectively.
- Apply troubleshooting techniques to provide solutions to our customer's individual needs.
- Troubleshoot, diagnose and potentially escalate customer inquiries during their engineering and operations efforts

PROFESSIONAL EXPERIENCE

HADOOP DEVELOPER

NYSE, New York City, NY **November 2016 - Present**

Responsibilities

- Hadoop installation, configuration of multiple nodes using Cloudera platform.
- Installed, configured and maintained Hortonworks HDP 2.2 using Ambari and manually through command line.
- Worked on analyzing Hadoop clusters using different big data analytic tools including Kafka, Pig, Hive and Map Reduce.
- Collected and aggregated large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
- Real time streaming the data using Spark with Kafka.
- Involved with Continuous Integration team to setup tool GitHub for scheduling automatic deployments of new/existing code in Production.
- Monitored multiple Hadoop clusters environments using Nagios. Monitored workload, job performance and capacity planning using MapR control systems
- Configured Spark streaming to receive real time data from the Kafka and store the stream data to HDFS using Scala.
- Worked within the Apache Hadoop framework, utilizing Opinion Lab statistics to ingest the data from a streaming application program interface (API), automated processes by creating Oozie workflows, and draw conclusions about consumer sentiment based on data patterns found through the use of Hive for external client use.
- Wrote the Storm topology with HDFS Bolt and Hive Bolts as destinations.
- Expertise in writing Storm topology development, maintenance and bug fixes.
- Developed Hadoop streaming Map/Reduce works using Java.
- Implemented test scripts to support test driven development and continuous integration.
- Worked on tuning the performance of Pig queries.
- Involved in loading data from Linux file system to HDFS and imported and exported data into HDFS using Sqoop.
- Good knowledge on building Apache spark applications using Scala.
- Experience working on processing unstructured data using Pig.
- Implemented Partitioning, Dynamic Partitions, Buckets in Hive.
- Implemented Spark using Scala and SparkSQL for faster testing and processing of data.
- Good knowledge with NoSQL databases like HBase, Cassandra
- Installed, administered, upgraded and managed distributions of Cassandra and involved in Cassandra performance tuning.
- Plan, deploy, monitor, and maintain Amazon AWS cloud infrastructure consisting of multiple EC2 nodes and VMWare VMs as required in the environment.
- Expertise in AWS data migration between database platforms like SQL Server to Amazon Aurora using RDS tool.

- Supported Map Reduce Programs those are running on the cluster.
- Managed and reviewed Hadoop log files.
- Involved in scheduling of Oozie workflow engine to run multiple pig jobs.
- Responsible for developing data pipeline using flume, Sqoop and Pig to extract the data from weblogs and store in HDFS.
- Data scrubbing and processing with Oozie.
- Developed Pig Latin scripts to extract data from the web server output files to load into HDFS.
- Involved in developing Hive DDLs to create, alter and drop tables.
- Created and maintained technical documentation for launching Hadoop clusters and for executing Hive queries and Pig Scripts.
- Responsible for Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, Data Frame, Pair RDDs, Storm, Spark YARN.
- Provided inputs on long term strategic vision and direction for the data delivery infrastructure including Microsoft BI stack implementations and Azure Advanced Data Analytic Solutions
- Evaluated existing data platform and apply technical expertise to create a data modernization roadmap and architect solutions to meet business and IT needs.
- Ensured technical feasibility of new projects and successful deployments, orchestrating key resources and infusing key data technologies (e.g. Azure Data technologies like, Azure Data Lake, Azure Blog Storage, Azure SQL DB, Analysis Services)
- Utilized Microsoft data bricks to process spark jobs and blob storage services to process data.
- Worked on data fabrics to process data silos of a big data system.
- Used Azure APP.FABRIC Message BUS to provide valuable functionality for integrating existing systems and building composite applications.
- Build a fully functional demo for Education Analytics Product using SharePoint Online, Windows Azure, SQL Azure, Silverlight technologies.
- Design and Implement Database Schema import data and build stored procedures on SQL Azure.
- Built Datasync job on Windows Azure to synchronize data from SQL 2012 databases to SQL Azure.
- Built SharePoint web-parts and Silverlight components for different dashboards.
- Maintained high degree of competency across the Microsoft Application Platform focusing on .NET Framework, WCF, Windows Azure, AppFabric, and SQL Azure.
- Leveraged broad and deep industry knowledge in architecting customer solutions by mapping business requirements to Azure based enterprise solutions.

- Built a prototype Azure application that accesses 3rd party data services via Web Services. The solution dynamically scales, automatically adding/removing cloud-based compute, storage and network resources based upon changing workloads.
- Built a consumer price comparison application on the Azure platform that requires no on-premises hardware. The application is designed to scale to meet the needs of an unpredictable fluctuating worldwide demand.
- Positioned Aflac to extend their on-premises computational functions to the Azure platform.
- Worked with Aflac infrastructure architects in resolving technical challenges.
- Worked directly with the Azure Product team in tracking and resolving defects in the beta Azure IaaS offering.
- Developed a conical map/reduce-like architectural pattern that that is designed to leverage the Azure platform. Work in progress on POC to migrate to Windows Azure to address scalability and performance issues.
- Make data fabrics to simplifies and integrate data management across cloud and on premises to accelerate digital transformation.
- Involved in Analyzing system failures, identifying root causes, and recommended course of actions. Documented the systems processes and procedures for future references.
- Worked with systems engineering team to plan and deploy new Hadoop environments and expand existing Hadoop clusters.
- Expertise in building Cloudera, Hortonworks Hadoop clusters on bare metal and Amazon EC2 cloud.
- Experienced in installation, configuration, troubleshooting and maintenance of Kafka & Spark clusters.
- Experience in setting up Kafka cluster on AWS EC2 Instances.
- Good understanding on cluster configurations and resource management using YARN
- Transform and analyze the data using PySpark HIVE, based on ETL mappings
- Application performance tuning to optimize resource and time utilization.
- Design application flow and implement end to end from gathering requirements, Build Code, perform testing, deploying into production
- Developed pyspark programs and created the data frames and worked on transformations.
- Worked spark transformations on source files to load the data into in hdfs.
- Developed performance tuning in spark program for different source systems domains and inserted into harmonized layer.
- Automated scripts using oozie and implement in production.
- Developed atomic scripts for scheduling oozie, Sqoop jobs daily or weekly basis.

Environment: Hadoop, HDFS, MAPREDUCE, HIVE, PIG, OOZIE, SQOOP, AMBARI, PYSPARK, STORM, GFS, ZOOKEEPER, NIFI, KAFKA, Microsoft Azure.

HADOOP DEVELOPER

Nationwide Insurance, Columbus, Ohio **January 2015 - September 2016**

Responsibilities

- Worked on analyzing Hadoop cluster and different big data analytic tools including Pig, Hive and Sqoop.
- Created POC on Hortonworks and suggested the best practice in terms HDP, HDF platform
- Set up Hortonworks Infrastructure from configuring clusters to Node
- Installed Ambari server on the clouds
- Perform architecture design, data modeling, and implementation of Big Data platform and analytic applications for the consumer products
- Worked as Administrator for Hadoop Cluster (180 nodes)
- Performed Requirement Analysis, Planning, Architecture Design and Installation of the Hadoop cluster
- Experience in Upgrades and Patches and Installation of Ecosystem Products through Ambari.
- Automated the configuration management for several servers using Chef and Puppet.
- Monitored job performances, file system/disk-space management, cluster & database connectivity, log files, management of backup/security and troubleshooting various user issues.
- Responsible for day-to-day activities which include HDFS support and maintenance, Cluster maintenance, creation/removal of nodes, Cluster Monitoring/Troubleshooting, Manage and review Hadoop log files, Backup restoring and capacity planning.
- Research and recommend suitable technology stack for Hadoop migration considering current enterprise architecture.
- Deployed Azure IaaS virtual machines (VMs) and Cloud services (PaaS role instances) into secure VNets and subnets.
- Provided high availability for IaaS VMs and PaaS role instances for access from other services in the VNet with Azure Internal Load Balancer.
- Script, debug and automate PowerShell scripts to reduce manual administration tasks and cloud deployments.
- Configure Implement, Secure and support Virtual Network and best security practices for single and multi-regional data centers.
- Created continuous integration system using Jenkins, Puppet & Chef full automation, Continuous Integration, faster and flawless deployments.
- Migrated the application from Infrastructure as a Service (IaaS) to Platform as a Service (PaaS) by converting existing solution to Windows Azure Worker Role.
- Developed Micro services tools using Python, Shell scripting, XML to automate some of the menial tasks.
- Moderate and contribute to the support forums specific to Azure Networking, Azure Virtual Machines, Azure Active Directory, Azure Storage) for Microsoft Developer Network including Partners and MVPs. expose, and automate utilizing custom data.

- Creation of business development offerings featuring all aspects of OMS architecture, deployment and solutions.
- Experience Microsoft Azure date storage and Azure Data Factory, Data Lake.
- Developed different kind of custom filters and handled pre-defined filters on HBase data using API.
- Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data.
- Configured Azure Traffic Manager to build routing for user traffic
- Infrastructure Migrations: Drive Operational efforts to migrate all legacy services to a fully Virtualized Infrastructure.
- Implemented HA deployment models with Azure Classic and Azure Resource Manager.
- Configured Azure Active Directory and managed users and groups
- Worked on tuning Hive and Pig to improve performance and solve performance related issues in Hive and Pig scripts with good understanding of Joins, Group and aggregation and how it does Map Reduce jobs
- Implemented concepts of Hadoop eco system such as YARN, MapReduce, HDFS, HBase, Zookeeper, Pig and Hive.
- In charge of installing, administering, and supporting Windows and Linux operating systems in an enterprise environment.
- Involved in Installing and configuring ranger for the authentication of users and Hadoop daemons.
- Experience in methodologies such as Agile, Scrum, and Test-driven development.
- Worked with cloud services like Amazon Web Services (AWS) and involved in ETL, Data Integration, Datawarehouse, and Migration, and installation on Kafka.
- Used Flume extensively in gathering and moving log data files from Application Servers to a central location in Hadoop Distributed File System (HDFS)Used Python and Django creating graphics, XML processing, data exchange and business logic
- Created Oozie workflows to run multiple MR, Hive and pig jobs.
- Supported in setting up QA environment and updating configurations for implementing scripts with Pig and Sqoop.
- Develop Spark code using Scala and Spark-SQL for faster testing and data processing
- Involved in the development of Spark Streaming application for one of the data sources using Scala, Spark by applying

Environment Hadoop, Microsoft Azure, Traffic Manager, HDFS, Pig, Sqoop, Shell Scripting, Ubuntu, Linux Red Hat, Spark, Scala, Hortonworks, Cloudera Manager, Apache Yarn, Python, Microsoft Azure.

HADOOP DEVELOPER

GAP, Groveport, Ohio **November 2013 - December 2014**

Responsibilities

- Evaluated business requirements and prepare detailed specifications that follow project guidelines required to develop written programs.
- Installed application on AWS EC2 instances and also configured the storage on S3 buckets.
- Performed S3 buckets creation, policies and also on the IAM role based polices and customizing the JSON template.
- Implemented and maintained the monitoring and alerting of production and corporate servers/storage using AWS Cloud watch.
- Installed, configured, monitored, and maintained HDFS, Yarn, HBase, Flume, Sqoop, Oozie, Pig, Hive.
- Worked on installing cluster, commissioning & decommissioning of Data Nodes, Name node recovery, Capacity planning, Cassandra and slots configuration.
- Experienced in developing programs in Spark using Python to compare the performance of Spark with Hive and SQL/Oracle.
- Hands on experience in provisioning and managing multi-node Hadoop Clusters on public cloud environment Amazon Web Services (AWS) - EC2 and on private cloud infrastructure.
- Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using Python. Worked with business teams and created Hive queries for ad hoc access.
- Experience in understanding Hadoop multiple data processing engines such as interactive SQL, real time streaming, data science and batch processing to handle data stored in a single platform in Yarn
- Hands on experience in installing, configuring MapR, Hortonworks clusters and installed Hadoop ecosystem components like Hadoop, Pig, Hive, HBase, Sqoop, Kafka, Oozie, Flume, Zookeeper.
- Monitoring systems and services, architecture design and implementation of Hadoop deployment, configuration management, backup, and disaster recovery systems and procedures.
- Worked on Scripting Hadoop package installation and configuration to support fully-automated deployments
- Supported Hadoop developers and assisting in optimization of map reduce jobs, Pig Latin scripts, Hive Scripts and HBase ingest required.
- Defined job flows and managed and reviewed Hadoop and HBase log files.
- Ran Hadoop streaming jobs to process terabytes of text data.
- Worked on YARN capacity scheduler by creating queues to allocate resource guarantee to specific groups.
- Implemented Hadoop stack and different bigdata analytic tools, migration from different databases to Hadoop (Hdfs)
- Configured Azure Traffic Manager to build routing for user traffic
- Infrastructure Migrations: Drive Operational efforts to migrate all legacy services to a fully Virtualized Infrastructure.

- Implemented HA deployment models with Azure Classic and Azure Resource Manager.
- Configured Azure Active Directory and managed users and groups
- Worked on tuning Hive and Pig to improve performance and solve performance related issues in Hive and Pig scripts with good understanding of Joins, Group and aggregation and how it does Map Reduce jobs
- Involved in scheduling Oozie workflow engine to run multiple Hive and Pig job
- Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring and troubleshooting, manage and review data backups, manage and review Hadoop log files.
- Managed datasets using Panda data frames and MySQL, queried MYSQL database queries from Python using Python-MySQL connector MySQL dB package to retrieve information.
- Developed various algorithms for generating several data patterns. Used JIRA for bug tracking and issue tracking.
- Developed backup policies for Hadoop systems and action plans for network failure.
- Involved in the User/Group Management in Hadoop with AD/LDAP integration.
- Resource management and load management using capacity scheduling and appending changes according to requirements.
- Implemented strategy to upgrade entire cluster nodes OS from RHEL5 to RHEL6 and ensured cluster remains up and running.
- Developed scripts in shell and python to automate lot of day to day admin activities.
- Installed several projects on Hadoop servers and configured each project to run jobs and scripts successfully
- Created user accounts and given users the access to the Hadoop cluster.
- Resolved tickets submitted by users, troubleshot the error documenting and resolved the errors.
- Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing.
- Implemented Spark using Scala and Sparks for faster testing and processing of data.

Environment: Hadoop, HDFS, Map Reduce, Hive, Pig, Puppet, Zookeeper, HBase, Flume, Ganglia, Sqoop, Linux, CentOS, Ambari, Microsoft Azure, Blob Storage

Selenium Tester

Ernst & Young, Chicago, IL **September 2010 - May 2013**

Responsibilities

- Responsible for implementation and ongoing administration of Hadoop infrastructure and setting up infrastructure
- Developed Test Plans, built test cases and test data sets based upon documented system requirements.
- Provided input on automating nightly builds, integration and regression, as well as exploratory and acceptance testing.
- Participated in process improvement and quality control activities.
- Review and understand business requirements/use cases and functional details
- Analyzed and implemented testing methods and equipment that aided to increase team confidence in the system.
- Evaluated test results and recommended changes in procedures to validate implementation against requirements.
- Participating in daily standups, Sprint planning, retrospective and grooming sessions.
- Conducting ATDD sessions with developers, UAT testers and product owner.
- Giving Demos of new features to PO and Stakeholders at the end of each Sprint.
- Analyzed and selected the test cases for automation of Java and Web application
- Performing manual testing of features within each sprint and automate features from previous Sprint.
- Created frameworks using TestNG and Web driver and parameterized the test for multiple sets of data testing
- Followed Agile Scrum methodology and waterfall methodology
- Arranged test suites to be able to upgrade tests easily in the event any feature or configuration changes.
- Modify the existing test cases based on change in a feature and requirements.
- Using Jira as a defect tracking tool for Product backlog and reporting bugs.
- Tested mobile UIs on iPhone, iPad, Android, BlackBerry, windows and other smart phones. Experienced in cross platform web testing on web browsers, iOS and Android devices.
- Created automation script using Selenium WebDriver using Java, Junit, TestNG XPath, CSS, Firebug and FirePath
- Responsible for mobile application automation using Appium and Java, Android Driver, Virtual Device Simulator.
- Experience in Dara Driving from excel for feeding data into Appium Testcases.
- Working on Android and iOS Automation Tools (Selenium, and Appium) for testing Native apps • Documenting test scenarios and test cases in a test case management system.
- Assisting UAT testers with data setup and execute business scenarios and wrote advanced SQL queries.

Environment: HP ALM, Selenium WebDriver, JUnit, Cucumber, Angular JS, Node.JS Jenkins, GitHub, Quality Center, Win Runner, LoadRunner, QTP, SQL Server 2000, VB.net

EDUCATION

Bachelor's in Business Administration,

Daffodil International University, Bangladesh

GLENN KHAN

- Queens, NY, USA

Contact Information

- gbigdata1983@gmail.com (Preferred)
- 9294356965 (Preferred)

Work History

Total Work Experience: 9 years

- **NYSE | HADOOP**
  Nov 01, 2016 - No End Date | New York City NY United States
- **| Nationwide Insurance**
  Jan 01, 2015 - Sep 01, 2016 | Columbus Ohio United States
- **GAP | HADOOP**
  Nov 01, 2013 - Dec 01, 2014 | Groveport Ohio United States
- **Selenium Tester | Ernst & Young**
  Sep 01, 2010 - May 01, 2013 | Chicago IL United States

Skills

- **infrastructure** | 8yrs | 2019
- **qa** | 8yrs | 2019
- **sql** | 8yrs | 2019
- **apache hadoop** | 8yrs | 2019
- **configuration** | 8yrs | 2019
- **software** | 8yrs | 2019
- **upgrades** | 8yrs | 2019
- **business requirements** | 7yrs | 2019
- **integration** | 7yrs | 2019

- **procedure** | 7yrs | 2019
- **linux** | 6yrs | 2019
- **mapreduce** | 6yrs | 2019
- **resource management** | 6yrs | 2019
- **scripting** | 6yrs | 2019
- **workflow** | 6yrs | 2019
- **amazon redshift** | 6yrs | 2019
- **apache flume** | 6yrs | 2019
- **apache hbase** | 6yrs | 2019
- **apache kafka** | 6yrs | 2019
- **apache oozie** | 6yrs | 2019
- **installation** | 6yrs | 2019
- **microsoft windows azure** | 6yrs | 2019
- **amazon web services** | 4yrs | 2019
- **agile** | 6yrs | 2018
- **api** | 6yrs | 2018
- **big data** | 6yrs | 2018
- **dba** | 6yrs | 2017
- **apache hive** | 4yrs | 2017
- **devops** | 6yrs | 2016
- **akka** | 4yrs | 2016
- **apache**
- **hadoop**
- **hbase**
- **hive**
- **microsoft windows**

Work Preferences

- Likely to Switch: Most Likely
- Willing to Relocate: No
- Work Authorization:
    - US
- Work Documents:
    - US Citizenship
- Desired Hourly Rate: 72+ (USD)
- Desired Salary: 145000+ (USD)
- Security Clearance: Yes
- Third Party: Yes
- Employment Type:
    - Contract - Corp-to-Corp
    - Contract - W2

    - Contract to Hire - Independent
    - Full-time
    - Contract to Hire - W2
    - Part-time
    - Contract - Independent
    - Contract to Hire - Corp-to-Corp

Profile Sources

- Dice: https://www.dice.com/employer/talent/profile/07d1b628e1eb77d250c700e81e3925e07c937fd8