**** Catherine Xu

cathyxu324@gmail.com (visa:H1B)

* **Around 7+ years** experience on **Java EE** development, 4+ years hands-on data experience on **Big Data Technologies** on **Hadoop** **ecosystems**
* Deep understanding/knowledge of **Hadoop Architecture** and major components such as **HDFS**, **MapReduce V1, YARN** architecture and good understanding of scalability, workload management, schedulers, and distributed platform architectures
* Good understanding of **microService** such as **jenkins**, **yaml, docker, kubernetes,** configuration, integration test and deployment
* Technical expertise in **Hadoop Map Reduce, Amazon EMR, HDFS**, **Hbase,** **Hive, Cloudera Manager,** **Zookeeper,** **Pig**, **Sqoop**, **Impala, SQL,** **Oracle SQL, MongoDB,** and also **Linux/UNIX Shell Scripting**.
* Experience in developing **MapReduce** jobs with **Java** **8, Python** **2.7/3.x** in **Hadoop**
* Experience in developing **Spark** applications using **Scala** for **Sparkstreaming** and **SparkSQL**
* Extensive experience in importing and exporting data using **Sqoop** from **HDFS/Hive/HBase** to Relational Database Management Systems (**RDBMS**) and vice versa.
* Experience in aggregating and moving large streaming data using **Flume, Kafka, RabbitMQ**
* Extensive experience in writing **Storm** and **Hive**/**Impala** Queries for processing and analyzing
* Experience in designing both time driven and data driven automated workflows using **Oozie**
* Experience on cluster security and authentication with **Kerberos**
* Experience on **MLib** and **Machine Learning** for data analysis
* Hands on experience with **data mining**, **machine learning** and underlying **algorithms**
* Strong understanding of **core Java**, **data structure**, **algorithm design**, **Object-Oriented Design** (**OOD**) **and Object-Oriented Programming (OOP),** **MVC**
* Hands on Java **Collections** **Framework**, **Exception** **handling** and **concurrency programming**
* Hands on experience in **Amazon Web Services (AWS EC2, EMR)**
* Hands on experience in **HTML, CSS, JSP, Jquery, Bootstrap** and Java frameworks like **Spring**
* Experience on **Apache Tomcat** and **Java Servlets**
* Experienced in **Tableau** Desktop for data visualization and analysis
* Experience on Java unit tests (**JUnit**) and **integration tests**
* Hands on flow chart & diagram software **Microsoft Visio**, **illustrator** and **Photoshop**
* Experience with source code management tools such as **Git Bash**, **GitHub**, **SVN, Bitbucket**
* Experience with project tracking software such as **Jenkins**
* Experience in **Agile** Development environments such as **Rally, itrack**
* **Hardworking** professional with a strong ability to work well in a team environment. Exceptional time management skills with a strong work ethic.

**CERTIFICATES:**

* CCA Spark and Hadoop Developer Exam (CCA175)
* Oracle Certified Professional, Java SE 8 Programmer
* Oracle Certified Associate, Java SE 8 Programmer

**TECHNICAL SKILLS:**

**Apache Hadoop Eco-system**: **Relational Databases**:

HDFS, MapReduce, YARN, Hive 0.13+,Pig, Oracle11g/10g/PostgreSQL 8.0,

Sqoop1.4.5, ZooKeeper, flume1,4+, EMR, impala, MySQL 5.x, Microsoft SQL Server 9.0

Kafka0.8.2+, RabbitMQ, Spark 1.4+, Oozie, MRUnit

**NoSQL Databases**: **Scripting**:

Hbase0.8, Cassandra2.1, MongoDB2.4+ UNIX Shell, HTML, XML, CSS, JSP, SQL

**Languages**: **Operation System**:

Java, Python, Scala, SQL, HiveQL, Pig Latin Ubuntu, CentOS, Windows, Mac OS

**Statics and Machinglearning Tools: Cloud Platform:**

SparkR, Mahout, MLlib Amazon Web Service, EMR, Heroku

**IDE Application:**

Sublime Text, Eclipse, PyCharm, Enthought Canopy,

Notepad++

**Others**:

microService(jenkins, yaml, docker, kubernetes, integration test and deployment)

Springmvc, REST, SOAP

Machine Learning

VMWare software, Vagrant, VirtualBox, XAMPP,

LAMP, Git, GitHub, SVN, Bitbucket, AWS,

Splunk, Mongoose, iPython(Jupyter),

**Professional Experience:**

**AT&T, Middletown, NJ May/17 – Now**

**Big Data Developer**

**Project: Big Data Common Machine Learning Platform**

AT&T Inc. is the largest provider of fixed telephony in the United States, and is also a provider of broadband and subscription television services. It is also the 18th-largest mobile telecom operator in the world with 135 million mobile customers.

The aim of the CMLP is using microservice architecture to implement a machine learning platform on big data processing, model file management and model prediction. Our team is responsible for extracting, transforming and loading modem/transportation data, creating microservices such as natural language processing service, model-manager service and datasource service.

**Responsibilities:**

* Developed and optimized **Spark Scala** programs/ **HiveQL** to perform data enrichment, tranformation and wrangling
* Implemented and supported in big data **ETL** procedure using **Scala/Spark Streaming**, **Python, presto** and other **Big Data tools**
* Used a **microservice** architecture, with a Spring-based service interacting through a combination of **REST** and **RDBMS** datasources.
* Build and maintain **SQL** scripts and complex queries for **Hive** data extraction and analysis
* Impelment **AIML** interpreter services and pattern recoginnition services to createcmlp platform inteligent Q&Aassistant **using Java**, **Lucene** and **OpenNLP**
* Responsible for creating model management microservice using **Spring Boot**, **MongoDB**, **Jgit**, **Postman** and **Swagger** to manage **PMML** files and model features.
* Create **keyword parser**, **SentenceDetector**, **Tagger** and **TextFeatureExtrator** in common-NLP library
* Experience on configure **configure-map**, **yaml**, **docker** and **Kubernetes deployment**
* Wrote **Junit tests** and **Integration test** cases for those microservices.
* **Onboarding training** for new joiners like **environment installation, access, service running, Postman, jenkins job, deployment** and **git turtorial**.
* Worked with **itrack** and **git bash**/**pull request** for project tracking
* Troubleshooting and optimization of performance, such as debug or optimize the code for better performance.

**Environment:**

**MicroService, Spring, MongoDB, HDFS, Spark, Hive, SQL, REST, Postman, Swagger, Jgit, NLP library, Lucene, Scala, Python, kubernete, docker, Jenkins, yaml, integration test, J2EE, Git Bash and itrack**

**AT&T, Middletown, NJ Oct /16 – April/2017**

**Analytic Software Developer**

**Project: Customer-Premises Equipment Advisor ( CPE advisor)**

AT&T Inc. is the largest provider of fixed telephony in the United States, and is also a provider of broadband and subscription television services. It is also the 18th-largest mobile telecom operator in the world with 135 million mobile customers.

The aim of the CPE advisor was to predict AT&T DSL Modem failure by extracting and analyzing AT&T user modems’ wifi log. Our team was focusing on writing data importing/exporting, data pre-processing, processing, sampling and modeling.

**Responsibilities:**

* Implemented and supported in big data **ETL** procedure using **Python**, **Sqoop**, **Pig** and **Big Data APIs**
* Enhanced data validation by using Python, **Beeline** and **Oracle ODBC**
* Experience on **HiveQL**and **Beeline** to run analytical functions, including **UDF** and **optimization**.
* Wrote script for auto data extraction and validation by using **Python script**
* Wrote **Python script** for sampling and sampling test
* Modified data modeling by using **MLlib** for machine learning purpose
* Hands on experience on **Pyspark**/ **Spark Scala** for data aggregation and enrichment
* Utilized **PyUnit, nose and pytest** for python tests
* Performed analysis and presented results using **SQL** and **PLSQL**
* Experience on **Apache Tomcat**, **Java Servlets** and **AWS s3**
* Implemented design patterns like **Data Access Objects (DAO),** Value Objects/Data Transfer Objects (**DTO**), **Singleton** etc
* Involved in design and development of presentation layer using **HTML** and **JSP**
* Worked with **Git Bash** for version control and **Rally** for project tracking, **Confluence** for documentation collaboration
* Troubleshooting and optimization of performance

**Environment:**

**Ambari (Hortonworks), Hadoop, YARN, HDFS, Spark, PySpark, MapReduce, Hive, Sqoop, Beeline, Oracle database 11g, Vagrant, Oracle SQL Developer, SQL, PLSQL, Python, Scala, PyUnit, J2EE, Tomcat, Git Bash and Rally**

**Western Union, San Francisco, CA Oct /15 – Oct/16**

**Hadoop Engineer**

**Project: Trusted Platform Modules (TPM)**

Western Union is a global financial services company offering consumer and business services such as money transfers, bill payments and prepaid debit cards. They have received over 508 million business payments in 2015 worldwide across 200 countries and territories.

The aim of the project was to analyze a huge stream of data by sorting customers’ information to stop fraud in wire transfer. Hadoop ecosystem was used to solve the fraud detection problem and graph fraud hot spots from flow of transaction. Our team was responsible for big data platform implementation, data query, log files reviewing, data analysis and graphing.

**Responsibilities:**

* Familiar with **Big Data architecture design** and **configuration**
* Installed and configured **Hadoop** **HDFS**, developed **Map Reduce** jobs in java for data cleaning
* Experienceon configuration of **Hadoop 2.0, Mapreduce v2.6.x**
* Experience on **Zookeeper** for cluster coordination services and **Kafka**
* Used **Flume** and **Kafka** to transform, enrich and stream transactions to different location
* Used **Sqoop** **job** to transfer structured and unstructured data from different resources
* Experience on **SQL** query and **SQL optimization**
* **NoSQL experience** on connecting and retrieving data from **Cassandra**
* Experience on creating **Hive** tables to store the processed results in a **tabular** format
* **Optimizing Hive tables** using **optimization techniques**to provide better performance with **HiveQL**
* Hands on experience of **importing data into RDD and aggregating statistics by Spark SQL/SparkStreaming** with **Scala**
* **Tableau** **9.2** experience on graphing results for reporting
* Involved in designing user screens and validations using **Core Java, HTML**, **jQuery**, **Ext JS** and **JSP**
* Developed Unit Test Cases, and used **JUNIT** for unit testing of the application
* Good understanding of **XML** methodologies (XML, XSL, XSD) including **Web Services**
* Experience on **Splunk** platform to review and monitor server log files
* Involved in troubleshooting and optimization of functional and performance

**Environment:**

**Zookeeper, Hadoop, YARN, SparkSQL, MapReduce v2.6x, Hive, Oracle database 11g, Hbase 0.9, Cassandra, Tableau 9.2, Splunk dashboard, kafka, Sparkstreaming, Python, J2EE, Scala, Maven**

**Thomson Reuters, Ann Arbor, MI Sept/14 – Sept/15**

**Software Engineer**

**Project: Customer Relationship Management (CRM)**

Thomson Reuters Corporation is a major multinational mass media and information firm operating in more than 100 countries, and has more than 60,000 employees around the world.

The project is by using Hadoop platform to analyze customer’s information such investment preferences, account lift cycle, average trading volume to find more valuable or potential customers and improve financial services. Our team was focusing on converting and cleaning data source files, processing data and collecting data for pattern model design.

**Responsibilities:**

* Designed and implemented **Hadoop 1.x** and **Map Reduce** based large-scale parallel processing
* Continuous monitoring and managing the Hadoop cluster through **Cloudera Manager**
* Writing **Sqoop** scripts to make the interaction between databases
* Experience on Hadoop analytical tools (**SparkR, Mahout, MLib)**
* Experience on **Oracle SQL** Database and **SQL** query.
* Relational Databases Experience such as **Teradata**, **Vertica** and **DB2**
* Experience on **NoSQL database** such as **HBase**
* Wrote **Hive queries** for data analysis and **query** **optimization**
* Used **Flume** **1.x** to stream and analyze the log data
* Data query on **HDFS** and **HBase** by using **Impala 1.x**
* Designed technical solution for real-time analytics using **Kafka** and Hbase
* To set up standards and processes for Hadoop based application design and implementation
* **Tableau** **8** experience on graphing results for reporting
* Developed and tested backend components and UI using **Java**, **JSP**, **EJB**, **HTML**, **DHTML**, **JavaScript**, and **XML**
* Working in an **Agile** team and **Git** for version contral
* Reviewing and monitoring data process log files

**Environment:**

**Hadoop 1.x, Map reduce, Cloudera Manager, HDFS, Sqoop, Hive, Oracle SQL Database, Teradata, Vertica, DB2, Hive, Flume, Hbase, Kafka, SparkR, Mahout, MLib, Tableau 8, Java**, **JSP**, **EJB**, **HTML**, **JavaScript**

**General Motors, Detroit, MI Feb/ 13 – Aug/2014**

**Database and Software Developer**

**Project: Dealership Performance Management**

**General Motors Company, commonly known as GM, is an American multinational corporation headquartered in Detroit, Michigan, which provides manufactures, markets, and distributes vehicles and vehicle parts, and sells financial services.**

The aim of the Dealership Performance Management project is **by using combined Geographic Information Systems and big data analytics to improve the performance of dealers, helping boost sales effectiveness. Our team was responsible to using Java API and Hadoop ecosystem to analyze data, interface implementation and code testing.**

**Responsibilities:**

* Experience includes **Google Map API, RESTful services** and **Core Java**
* Implemented the system using **J2EE, Maven, Spring MVC** and **Eclipse**, **Apache** **Tomcat**
* Code testing by using **JAVA** **JUnit test and Integration test**
* Inside SQL Server databasemaintenance
* Migrate current scale **Microsoft SQL Server** to **Oracle Database**
* Collaborate with teammates to set up the sample databasefor scale test
* Set up the testing environment through VMware software for **Scale Testing**
* Experience on setting up and configuration of Amazon **EMR** on **AWS**
* Configured and optimized **Hadoop 1.x** and **Map Reduce V1** environment
* Implemented (Extract, Transform, Load) **ETL** to load data into **HDFS** with **Sqoop1.4**
* Handled data transformations using **Hive** and loaded data into HDFS
* Analyzed large amounts of data sets to determine optimal way to aggregate and report on
* Experience on using **Oozie** to manage Apache Hadoop job
* Extending **HIVE** and **Pig 0.1** core functionality by using **Hcatalog**
* Working in an **Agile** team and **Git/Jenkins** for version control

**Environment:**

**Google Map API,** **RESTful, Apache Tomcat, JUnit Hadoop 1.x, Microsoft SQL Server, Oracle Database**, **MapReduce v1, Hive0.1, Pig 0.1, Hcatalog, Scoop 1.4, HBase 0.9, HDFS, Oozie, Amazon EMR, Amazon AWS, core Java, Maven, Spring MVC** and **Apache** **Tomcat**

**Quicken Loans Inc.** **Detroit, MI Jan/12–Jan/13**

**Software Developer**

**Project: Data Storage Modeling**

Quicken Loans Inc. is one of the largest full-service residential mortgage lenders located in the heart of the financial district of downtown Detroit, Michigan.

This project was developing the interface to track data storage from various sources. Our team is responsible for code design, interface and storage database connection implementation and code review/testing.

**Responsibilities:**

* Using **OOP/OOD** designed point of storage interface based on **J2EE** and **Netbean**
* Understanding of database design/architecture and experience on **MS SQL**/**RDBMS/DB2** databases
* Strong understanding of **IIS**, **XML (DTD, SOAP), JSON**, **XSL, WSDL**
* Familiar with large data models
* Implemented the system using **Maven,** **Eclipse**, **Log4j, Apache** **Tomcat**
* Extensive use of **HTML**, **CSS** and **JQuery** for the presentation layer along with **JavaScript** for the client side validations.
* Developed the UI screens using **JSP** and **JSF**
* Working in an **Agile** team and team up with QA
* Experience with project tracking software such as **Jira**
* Experience on version control tool **Git**

**Environment:**

**J2EE, Log4j, Apache Tomcat, SQL/RDBMS/DB2, XML, DTD, SOAP, WSDL, JSON, OOP/OOD, Shell Scripting, Linux, Maven, HTML, CSS, JQuery, JSF, JSP**

**NetEase, Beijing, China Aug/09 - Nov/11**

**Java Developer**

**Project: Online Payment System and advertisement recommendation**

NetEase, Inc. is a key pioneer Chinese Internet technology company providing online services centered on content, community, communications and commerce.

The aim of the project was to develop the new platform of online payment system on the online e-Commence services. By using the third money transfers party, Yeepay, our team was responsible for the design, coding and testing the payment interface, authentication and testing.

**Responsibilities:**

* Used **HTML, CSS, JSP** to implement the payment interface based on **Java 6** and **Eclipse**
* Designed the MySQL database to handle the large amount of data easily by using relational database schema, table partitioning, etc
* Familiar with **MySQL** data query and **JDBC** for database connectivity
* Experience on **REST-ful Web Services**
* Implementation information authentication with **cryptographic algorithm**
* Experience on **Apache Tomcat** and **Java Servlets**
* Experience on unit tests (**JUnit**), **integration tests** and configuration changes
* Used **Log4j** for External Configuration Files and debugging.
* Working with QA team for finalizing Test Plans.
* Worked in **Agile** team

**Environment:**

**MySQL, JUnit, J2EE, Eclipse, Apache Tomcat 6, HTML, CSS, JSP, JDBC, JavaScript, Agile**

**EDUCATION:**

**Master of Science, Engineering in Computer Science**