**Clinton todd**

Clinton.todd09@outlook.com

**KEY AREAS OF EXPERTISE**

* Around 8 years of IT experience as **Associate Java Developer** & **Big Data/Hadoop Developer**.
* Passionate towards working in Hadoop & Big Data Architectures, Big Data Processing, Analysis & Visualization.
* Good knowledge of Hadoop Distributed File System & Ecosystem components like **MapReduce, Hive, Pig, Hbase, Zookeeper, Flume, Splunk, Sqoop, Storm, Kafka, Oozie & Spark streaming, Core Spark API**.
* Good understanding of Hadoop **MapR** distributions.
* Ecosystems familiarity with Cloudera **CDH1, CHD2, CDH3, CDH4, CDH5**.
* Execution of Batch jobs through the data streams through **SPARK Streaming**.
* Experience in Apache **Spark**, **Spark** **Streaming**, **Spark SQL** and **No SQL** databases like **Cassandra** and **Hbase**.
* Have experience in Hadoop distributions like **MapR,** **AWS, Cloudera**.
* Have thorough knowledge on **spark architecture** and how **RDD's** work internally. Have exposure to **Spark Streaming** and **Spark SQL**.
* Integrated Spark Streaming with Sprinkler using pull mechanism and loaded the **JSON** data from social media into **HDFS system**.
* Have experience in Shell Scripting like **Scala/Python scripting languages** and used it extensively with **Spark** for data processing.
* Detailed understanding of Hadoop internal architecture and functionality of various components such as **Job Tracker, Task Tracker, Name Node & Data Node, Application Master, Resource Manager, Node Manager & MapReduce** programming paradigm.
* Developing various cross platform products while working with different Hadoop file formats like Sequence File, **RC File, ORC, AVRO & Parquet**.
* Analyzing Data through **Hive QL, Pig Latin & MapReduce** programs in Java. Extending HIVE and PIG core functionalities by implementing **custom UDF’s**.
* Installing and setting up Hadoop environment in cloud through the **AMZON WEB SERVIECS**.
* Implementing, **partitioning and bucketing in Hive** for more efficient querying of data.
* Extensively involved in design, development, tuning and maintenance of **HBase** Casandra databases.
* Extensive expertise in creating and Automation(Maven) of workflows using **Oozie** workflow Engine.
* Scheduled jobs using **Oozie Coordinator**, to execute jobs on specific days (excluding weekends).
* Hands on experience on **Apache, Cloudera & MapR** Hadoop environments.
* Good understanding of Hadoop administration with Cloudera & MapR.
* Experience in using, Jenkins, **Maven 2.0** to compile, package and deploy to the **application servers**.
* Great expertise experience in Requirements gathering, Design, Development & Reporting applications.
* Expertise in data modelling & deployment strategies in production environment meeting **Agile requirements**.
* Proficient in coding and optimizing **Teradata** batch processing scripts for data transformation, aggregation and load using BTEQ.
* **Teradata** consultant with experience in **Teradata** Physical implementation and Database Tuning
* Strong experience in coding and debugging **Teradata** utilities like Fast Load, Multiplied and Tpump for loading data from various sources into Teradata and data export using Fast Export.
* Release Automation while trouble shooting. (**CA Automation, Jenkins, Ant & Maven build** & **IBM SCM** Architectures)
* Experience in creating functional flow, technical design specification document based on the functional requirement adopting **Agile Methodologies** involving code check-in, checkout from repositories.
* Proficiency in working with **PL/SQL** implementation on **Data warehousing** concepts and strong experience in implementing data warehousing methodologies.
* Experienced in using waterfall, **Agile and Scrum models** of software development process framework.
* Experienced in using Jenkins & **Maven 3.3.** to compile the package and deploy to the Application Servers.
* Deployment, Distributed and Implementation of Enterprise applications in **J2EE** environment.
* Comprehensive knowledge of **Software Development Life Cycle (SDLC)**, having thorough understanding of various phases like Requirements Analysis, Design, Development and Testing.
* Experienced with different modules of spring like spring - **AOP, IOC/Core & MVC**.
* Good understanding of **bootstrap, spring rest and integration**.
* Extensive hands-on experience with core expertise in design, development and deployment of N-Tier enterprise applications for J2EE platform using Java, Java Script, Struts, Spring, EJB, Servlets, JSP, Web services, **Angular JS & Node JS,** **JNDI, JSON JMS, JAXP, JUnit and XML**.
* Proficient in developing web page quickly and effectively using, **HTML 5, CSS3, JavaScript and jQuery** and also experience in making web page cross browser compatible.
* Worked on generating several reporting tools in collaboration to reporting tools like **Talend & Tableau**
* Experience in developing applications with **Angular JS & Node JS**.
* Experienced with Java Multithreaded programming to develop multithreaded modules and applications, Experience in Development of **Multi-Tier distributed application** using Java and Technologies.
* Experience in monitoring, troubleshooting and supporting J2EE based applications and infrastructure.
* Proven experiences of using Application Servers like **Web Sphere, Tomcat, Web Logic, JSON, JBoss, Tomcat**.
* Good understanding of RDBMS which includes writing queries, stored procedures using Oracle […] MS SQL Server, AS400 and DB2.
* Strong experience in **Unix Scripting** and running **Python** & **Scala** Scripts.
* Good Working knowledge of complex workflow schedulers like **Oozie** & UC4.
* Implementing database driven applications in Java using **JDBC, JSON, XML API** and using hibernate framework.
* Proficient in using Design Patterns like GOF Design patterns (Creational, Structural and behavioral) and J2EE design patterns like **MVC, Singleton, Front Controller**, Business Delegate, Service Locator, **DAO, VO, DTO** etc.
* Involved in the Software Life Cycle phases like **AGILE and Waterfall** estimating the timelines for projects.
* Experience in Installation, Configuration and Re-Configuration of **Oracle Database** on **Red hat Linux** **AS/ES/4/5** and Windows advanced servers and in **RAC environment**.
* Strong knowledge of version control systems like **SVN, CVS & GIT**.

**PROGRAMMING/TECHNICAL SKILLS:**

* **Big Data Skillset - Frameworks & Environments:**

Cloudera CDHs, Hortonworks HDPs, Hadoop1.0, Hadoop2.0, HDFS, MapReduce, Pig, Hive, Impala, HBase, Data Lake, Cassandra, MongoDB, Mahout, Sqoop, Oozie, Zookeeper, Flume, Splunk, Spark, Storm, Kafka, YARN, Falcon, Avro.

* **Amazon Web Services(AWS):**

Elastic Map Reduce(EMR), Amazon EC2, Amazon S3, [AWS CodeCommit](https://aws.amazon.com/documentation/codecommit/), [AWS CodeDeploy](https://aws.amazon.com/documentation/codedeploy/), [AWS CodePipeline](https://aws.amazon.com/documentation/codepipeline/), [Amazon CloudFront](https://aws.amazon.com/documentation/cloudfront/), [AWS Import/Export](https://aws.amazon.com/documentation/importexport/).

* **JAVA & J2EE Technologies:**

Core Java(Java8 & Java FX versions), Hibernate framework, Spring framework, JSP, Servlets, Java Beans, JDBC, EJB 3.0, Json, Java Sockets & Java Scripts. JavaScript, jQuery, JSF, Prime Faces, XML, […] Servlets, EJB, JDBC, HTML, XHTML, CSS, SOAP, XSLT and DHTML

Messaging Services JMS, MQ Series, MDB, J2EE MVC Frameworks Struts […] Struts 2.1, Spring 3.2, MVC, Spring Web Flow, AJAX.

* **IDE Tools:**

Eclipse, Net Beans, Spring Tool Suite, Hue (Cloudera specific).

* **Web services & Technologies:**

XML, HTML, XHTML, JNDI, HTML5, AJAX, jQuery, CSS, JavaScript, AngularJS, VB Script, WSDL, SOAP, JDBC, ODBC Architectures REST, MVC architecture.

* **Databases & Application Servers:**

Oracle, MySQL, DB2, Cassandra, MangoDB, Hbase, MangoDB, Database Technologies MySQL, Oracle 8i, 9i, 11i & 10g, MS Access, Teradata, Microsoft SQL-Server 2000 and DB2 8.x/9.x, PostgreSQL.

* **Other Tools:**

Putty, WinScp, DataLake, Talend, Tableau, GitHub, SVN, CVS.

**WORK EXPERIENCE**

**Role: Big Data Engineer Sept’2016 –Till date**

**Client: American Express, Phoenix, AZ.**

American Express Company, also known as Amex, is an American multinational financial services corporation headquartered in Manhattan's Three World Financial Center in New York City, United States. Founded in 1850, it is one of the 30 components of the Dow Jones Industrial Average. The company is best known for its credit card, charge card, and traveler's cheque businesses. Amex cards account for approximately 24% of the total dollar volume of credit card transactions in the US. American Express provides innovative payment, travel and expense management solutions for individuals and businesses of all sizes. Most of all, we help our customers realize their dreams and aspirations through industry-leading benefits, access to unique experiences, business-building insights, and global customer care. We enable our customers to do more and achieve more.

**Roles & Responsibilities:**

* Implemented several scheduled **Spark, Hive & Map Reduce jobs** in Hadoop **MapR** distribution.
* Cluster size is scalable and with Band width of 40GB nodes with a twostep process namely **Data Ingestion** flowed by **Data Processing**.
* Deployed several process oriented scheduled jobs through **cron tabs** and **UC4** **event engine work flow schedulers** using **wrapper scripts** for invoking the **Spark module**.
* Developed various main & service classes through **Scala** using **spark SQLs** for the requirement specific tasks.
* Expertise Understanding of Apache Spark High level architecture and cluster integration.
* Hands on coding with Scala for leveraging Apache Spark through the Scala APIs.
* Performed Data preprocessing and data cleaning using Hive & Pig.
* Project scope limited to Only Spark SQL and Spark Streaming here but interested in exploring Mlibs.
* Populating Hbase tables through automation. Used Spark Shell for querying database.
* Great familiarity with **Hive joins** & used **HQL** for querying the databases eventually leading to complex **Hive UDFs**.
* Veracity of Data from various sources: Sequence file, Avro, Text, Hive Tables, Batch Streams and Logfiles.
* **Wrapper scripts** in Unix Shell for automation using **shell scripting**.
* Build Tools used were **Maven and Sbt**.
* Familiarity with fixing the Build integration process and schema design changes
* Involved in **Data Validation** and fixing discrepancies by working in coordination with the **Data Integration** and Infra Teams.
* Productive implementation of **DStreams** on resilient distributed dataset (**RDD**) through various windows also simultaneously update log files for the streams.
* Extensive experience in Spark Streaming (version 1.5.2) through core Spark API running Scala, Java to transform raw data from several data sources into forming baseline data.
* Hands on expertise in running the **SPARK & SPARK SQL**.
* Implemented **SPARK batch jobs**.
* Developing the Tasks and setting up the requirement environment for running Hadoop in cloud on various instances.
* Developed **Hive (version 0.11.0.2)** and **Impala (2.1.0 & 1.3.1)** for end user / analyst requirements to perform ad hoc analysis.
* Automated complex work flow schedulers using the UC4 work flow scheduler.
* Performed Unit Testing & Integration with sample test cases and assisted QA Team and addressed several performance issues according to the Business Unit requirements.
* Proven expertise in handling the exception scenarios while handling the errored feed data in coordination with the Data Architects, Data Integration team, Business Partners and Stakeholders.
* Indulged in regular stand-up meetings, status calls, Business owner meetings with stake holders, Risk management Teams in an Agile Environment.
* Great understanding of the high-level architecture of the business logic for decomposing the complexity of module to simple achievable tasks for efficient development.
* Ideal approachability towards handling issues raised against the team by being a great team player as well as shouldering the responsibility as an individual in an appropriate manner whenever it matters.

**Environment:** MapR Hadoop Distribution, M3&M5, Hive, Scala, HBase, Sqoop, Maven builds, Spark, Spark SQL, Oozie, Linux/Unix, Shell Scripting, UC4 Complex workflow, SVN, Talend, Lambda, Akka, Kafka.

**Role: Hadoop Spark Developer May’2014 –Aug’16**

**Client: Hewlett Packard Enterprise, Houston, TX.**

Hewlett Packard has been in the innovation business for more than 75 years. We deliver high-quality, high-value products, consulting, and support services in a single package. We make IT environments more efficient, productive and secure, enabling fast, flexible responses to a rapidly changing competitive landscape. We enable organizations to act quickly on ideas by delivering infrastructure that can be easily composed and recomposed to meet shifting demands, some of our customers run traditional IT environments. Most are transitioning to a secure, cloud-enabled, mobile-friendly infrastructure. Many rely on a combination of both. Wherever they are in that journey, we provide the technology and solutions to help them succeed.

**Roles & Responsibilities:**

* + Worked on a live 60 nodes Hadoop cluster running **CDH5.4.4, CHD5.2.0, CDH5.2.1**
  + Responsible for migrating the code base from Cloudera Platform to other eco systems components like **Redshift, Dynamo DB**.
  + Involved in migrating the map reduce jobs into Spark Jobs and Used **Spark Transformations & Actions**  by leveraging the RDD data structures through scala..
  + Involved in requirement and design phase to implement Streaming Lambda Architecture to use real time streaming using **Spark and Kafka**.
  + Worked with highly unstructured & semi structured data of 90 TB (270 TB with replication factor of 3)
  + Extracted the data from Teradata and RDBMS into **HDFS**/Databases/Dashboards.
  + Worked on SPARK engine creating batch jobs with incremental load through **STORM,** **KAFKA, SPLUNK, FLUME, HDFS/S3, KINESIS, Sockets** etc.
  + Productive implementation of **DStreams** on resilient distributed dataset **(RDD)** through various windows also simultaneously update log files for the streams.
  + Extensive experience in **Spark Streaming** (version 1.5.2) through core **Spark API** running Python, Scala, Java & Python Scripts to transform raw data from several data sources into forming baseline data.
  + Hands on expertise in running the **SPARK & SPARK SQL**.
  + Implemented **SPARK** batch jobs.
  + Developing the Tasks and setting up the requirement environment for running Hadoop in cloud on various instances.
  + Developed **Hive (version 0.11.0.2)** and [**Impala (2.1.0**](#impala_downloads_210_unique_2) **&** [**1.3.1**](#impala_downloads_131_unique_2)**)** for end user / analyst requirements to perform ad hoc analysis
  + Implemented various **MapReduce Jobs** in custom environments and updating them to **Hbase tables** by generating hive queries.
  + Performed Sqooping for various file transfers through the Hbase tables for processing of data to several **NoSQL DBs.**
  + Involved in developing **Hive UDFs** and reused in some other requirements.
  + Worked on performing Join operations.
  + Involved in creating partitioning on external tables.
  + Good hands on experience in writing **HQL statements** as per the user requirements.
  + Implemented Cassandra connector for Spark in Java.
  + Implemented Cassandra connection with the **Resilient Distributed Datasets** (local and cloud)
  + Data visualization and reporting was performed through **Tableau & Talend**.
  + Very good understanding of **Partitions, Bucketing**concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
  + Solved performance issues in **Hive and Pig scripts** with understanding of **Joins, Group** and aggregation and how does it translate to **MapReduce jobs**.
  + Developed **UDFs in Scala, Java & Python** as and when necessary to use in **PIG and HIVE** queries
  + Developed **Oozie workflow** for scheduling and orchestrating the ETL process
  + Implemented authentication using **Kerberos** and authentication using Apache Sentry.
  + Worked with the admin team in designing and upgrading **CDH 4 to CDH 5**
  + Good working knowledge of Amazon Web Service components like **EC2, EMR, S3** etc.
  + Examining job fails and trouble shooting.
  + Very good experience in monitoring and managing the **Hadoop cluster using Cloudera Manager**.
  + Good working knowledge of **HBase**
  + Good Working knowledge of **Talend & Tableau**

**Environment:** Cloudera CHD4, Cassandra, Amazon Web Services, CDH5, Scala, Python Java Scripts Hive, Sqoop, Splunk, Storm, Spark, Flume,AVRO, Oozie, CentOS, Ambari, Oracle, SVN, Kafka, Data Lake, Github Java 1.7.x, JIRA, Talend.

**Role: Hadoop Developer Aug’12-April’14**

**Client: United Health Group, Boston**

United Health Group is an operating division of UnitedHealth Group, the largest single health carrier in the United States. At united Health Group, we are committed to improving the health care system, Improve the quality and effectiveness of health care for all Americans, enhance access to health benefits, create products and services that make health care more affordable, and Use technology to make the health care system easier to navigate.

**Roles & Responsibilities:**

* + Determining the viability of a business problem for a Big Data solution.
  + Worked on a **42 nodes** Hadoop Data Platform running **MapR Distribution.**
  + Worked with highly unstructured and semi structured data sets of 120 TB in size.
  + Extracted the data from **RDBMS into HDFS** using **Sqoop** throughIncremental Import.
  + Created and worked **Sqoop (version 1.4.3)** jobs with incremental load to populate Hive External tables.
  + Setting up nodes for performing MapReduce Jobs.
  + Extensive experience in writing **Pig (version 0.11)** scripts to transform raw data from several data sources into forming baseline data.
  + Developed **Hive (version 0.10)** scripts for end user / analyst requirements to perform predictive analysis
  + Implemented data access jobs through **Pig, Hive, Hbase (0.98.0), Storm (0.9.1)**
  + Solved performance issues in Hive and Pig scripts with understanding of Joins, Group and aggregation and how does it translate to MapReduce jobs.
  + Developed **Oozie** workflow for scheduling and orchestrating the process
  + Implemented authentication for **No SQL DB** through MangoDB & Cassandra connectors scripts for enabling the scripts to run.
  + Very good experience with both **MapReduce 1 (Job Tracker) and MapReduce 2 (YARN)** setups
  + Involved in selecting the right products to implement a Big Data solution.
  + Setup and Install Hadoop **MR1 cluster** and Enterprise Data Ware House.
  + Implemented analytics using **PIG & HIVE**.
  + Plan and manage **HDFS** storage capacity. Advise a team on best tool selection, best practices, and optimal processes using **Sqoop, Flume, Splunk, Pig, Oozie, Hive, and Bash Shell Scripting & Python Scripting**.
  + Facilitate access to large data sets utilizing **Pig/Hive** on Hadoop Ecosystem.
  + Fetching the **HQL results into CSV files** and handover to reporting team.
  + Good hands on experience in working with hive complex datatypes.
  + Imported millions of structured data from relational databases using **Sqoop import** to process using Spark and stored the data into **HDFS** in **CSV** format.
  + Developed mappings/workflows to transfer data from Oracle to **HDFS** and vice versa tables using **Hadoop\_hdfc\_connection**.
  + Developed mappings using data processor transformation to load data from word, pdf documents to HDFS.
  + Used parser to load structured and unstructured data.
  + Resolving the Context based scenarios for faster through put.
  + User defined functionality of the **Cassandra Architecture**.
  + Hands-on expertise with various architectures **Cassandra DB**.
  + Very good experience in monitoring and managing the **Hadoop cluster** using **Hortonworks**
  + Good working knowledge of **Cassandra Architecture**
  + Reporting Expertise through **Talend**

**Environment**: MapR Data platform, GitHub, RDBMS, Cassandra, Unix, Pig, Hive, Storm, Maven, Tableau, RDBMS, MS SQL VizualStudio, Talend.

**Role: Hadoop Developer. Aug 2010-July 2012**

**Client: State Farm, Bloomington, IL.**

State Farm provides insurance and financial services to their customers in the United States. The main purpose of this project is to store and analyze the data in Hadoop for real time decisions and to improve promotions to its customers.

* **Roles & Responsibilities:** 
  + Installed and configured Hadoop **MapReduce, HDFS**, developed multiple MapReduce jobs in Java for data cleaning and preprocessing.
  + Importing and exporting data into **HDFS from Oracle database** and vice versa using Sqoop.
  + Experience in installing, configuring Hadoop cluster for major **Hadoop distributions**.
  + Experience in using **Hive and Pig** as an ETL tool for event joins, filters, transformations and pre- aggregations.
  + Created **partitions, bucketing** across state in **Hive** to handle structured data.
  + Implemented Dash boards that handle **HiveQL** queries internally like Aggregation functions, basic hive operations, and different kind of **join operations**.
  + Implemented business logic based on state in **Hive using Generic UDF's**.
  + Developed workflow in **Oozie** to orchestrate a series of **Pig scripts** to cleanse data, such as removing personal information or merging many small files into a handful of very large, compressed files using Pig pipelines in the data preparation stage.
  + Experienced in writing the **Map Reduce** programs for analyzing of data as per the business requirements.
  + Used Pig in three distinct workloads like **pipelines**, **iterative** processing and research.
  + Involved in moving all log files generated from various sources to **HDFS** for further processing through **Kafka, Flume & SPLUNK** and process the files by using **Piggybank**.
  + Extensively used **PIG** to communicate with **Hive using HCatalog** and **HBASE** using **Handlers**.
  + Implemented **MapReduce jobs** to write data into **Avro format**.
  + Created Hive tables to store the processed results in a tabular format.
  + Exported the analyzed data to the **relational databases using Sqoop** for visualization and to generate reports for the BI team.
  + Implemented various **MapReduce** Jobs in custom environments and updating them to **Hbase** tables by generating **hive** queries.
  + Performed **Sqooping** for various file transfers through the **Hbase** tables for processing of data to several **NoSQL DBs- Cassandra, MangoDB**.
  + Involved in developing **Hive UDFs** and reused in some other requirements.
  + Worked on performing Join operations.
  + Involved in creating partitioning on external tables.
  + Experience in using Pentaho Data Integration tool for data integration, **OLAP** analysis and **ETL** process
  + Used **Hadoop's Pig, Hive and Map Reduce** for analyzing the Health insurance data and transforming into data sets of meaningful information such as medicines, diseases, symptoms, opinions, geographic region detail etc.
  + Worked on **data analytics** using Pig and Hive on Hadoop.
  + Evaluated **Oozie** for workflow orchestration in the **automation of MapReduce jobs, Pig and Hive jobs**.
  + Installed and configured **Hadoop Map Reduce, HDFS**, developed multiple Map Reduce jobs in java for data cleaning and preprocessing.
  + Created tables, secondary indexes, join indexes, viewed. in Teradata development Environment for testing.
  + Assisted in Batch processes using Fast Load, BTEQ, UNIX Shell and Teradata SQL to transfer, cleanup and summarize data.
  + Developed MLOADscripts to load data from Load Ready Files to **Teradata** Warehouse
  + Extracted files from MongoDB through **Sqoop** and placed in **HDFS** and processed
  + Captured the **data logs** from web server into **HDFS using Flume & Splunk** for analysis.
  + Experienced in writing **Pig scripts and Pig UDFs** to pre-process the data for analysis.
  + Experienced in managing and reviewing Hadoop log files.
  + Built front end using **JSP, JSON, Servlets, HTML** and JavaScript to create user friendly and appealing interface.
  + Reporting Expertise through **Talend.**
  + Used **JSTL** and built custom tags whenever necessary.
  + Used Expression Language to tie beans to **UI components**.
  + Gained very good business knowledge on health insurance, claim processing, fraud suspect identification, appeals process etc.

**Environment:** Hive, Pig, MapReduce, AVRO, Sqoop, Oozie, Flume,Kafka, Storm, HBase, Unix, Python, Sql, Hadoop 1.x, HDFS, Talend, Pig, Hive, Hbase, Github, MapReduce, Java, Sqoop, Flume, Splunk, Oozie, Linux, UNIX Shell & Python Scripting.