1. **LAMBDA ARCHITECTURE**

Lambda Architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch and stream processing methods. (WIKI: <https://en.wikipedia.org/wiki/Lambda_architecture>)

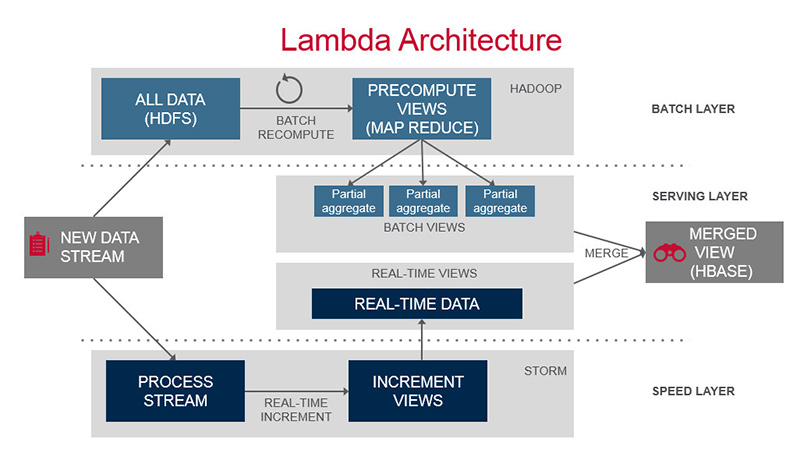
This design consists of three layers; **Batch**, **Speed**, and **Serving**.

1. **Batch** - A distributed processing system that can handle very large quantities of data. This is the layer where Apache Hadoop exists.

Functionality can also be described in terms of 1) Managing the master data set as an append only, immutable set of raw data, and 2) pre-computing arbitrary query functions called batch views.

1. **Speed** - This layer can process data streams in real time and without the need for completeness. Fills the gap caused by the batch layer’s lag in providing views based off of most recent data. Sacrifices throughput as it aims to minimize latency by providing real time views of most recent data. Technologies that exist in this layer include Apache Storm, SQLStream and Apache Spark.  
   Functionality can also be described as 1) Accommodating all requests that deal with low latency requirements. Uses fast algorithms dealing with recent data only.
2. **Serving** - Stores output from other layers. This layer responds to ad-hoc queries by returning pre-computed views or building new views from processed data. Apache HBase exists in this layer.

Functionality can also be described as 1) Indexing batch views so that they can be queried with low latency.



*Image provided by MapR @* [*http://www.mapr.com/sites/default/files/otherpageimages/lambda-architecture-2-800.jpg*](http://www.mapr.com/sites/default/files/otherpageimages/lambda-architecture-2-800.jpg)

1. **HDINSIGHT**

Azure HDInsight deploys and provisions managed Apache Hadoop clusters in the cloud, providing a software framework designed to process, analyze, and report on big data with high reliability and availability. HDInsight uses the Hortonworks Data Platform (HDP) Hadoop distribution. Hadoop often refers to the entire Hadoop ecosystem of components, which includes Apache HBase, Apache Spark, and Apache Storm, as well as other technologies under the Hadoop umbrella. (MS: <https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/>)

HDInsight can be accessed through the Azure Portal. Here clusters can be managed and created.

1. **Use Case: “Getting Users Near Me”**

Users are sending data into our system constantly. We can assume for our purposes their locations are captured by our system and saved with a time stamp. In the context of the lambda design pattern our system could behave in the following way.

1. All users are sending location data
2. User accesses this feature
3. Pre-computed views mixed with real-time data at the serving layer deliver a set of coordinates.
4. Device plots these points on a map
5. If this method proves too slow, access to partial set of data can be accessed quicker at the speed level. With left-out data to be delivered later.
6. **Testing In Java**
   1. **Tools**

* **Junit:** [**http://arquillian.org/**](http://arquillian.org/)
* **JTest:** [**https://www.parasoft.com/product/jtest/**](https://www.parasoft.com/product/jtest/)
* **Arguillian:** [**http://arquillian.org/**](http://arquillian.org/)
  1. **Integration Testing**

After Unit testing occurs, they must be tested to make sure they work together as expected. Integrate/combine the unit tested module one by one and test the behavior as a combined unit.

* 1. **Autograders**: Automated grading of Java software based off of defined project guidelines. Useful for creating standardized code across team.
  2. **ANT**: Java build tool to manage, compile, and package into JAR.
  3. **Unit Testing**: Individual modules testing independently and completely based on functionality and expected behavior.

1. **Hadoop**
   1. **What Needs to Go In?**: Apps and devices deliver data into the HDFS via HDInsight. Our MYSQL database can connect through

Tools:

* HDInsight Emulator (Windows Only, HDBase and Storm not included)
* Hadoop Sandbox (Same as Above for Linux/OSX)
* JDK 7 +
* Apache Maven
* Azure Subscription
  1. **What Needs to Come Out?**
* Query results
* Analytics results

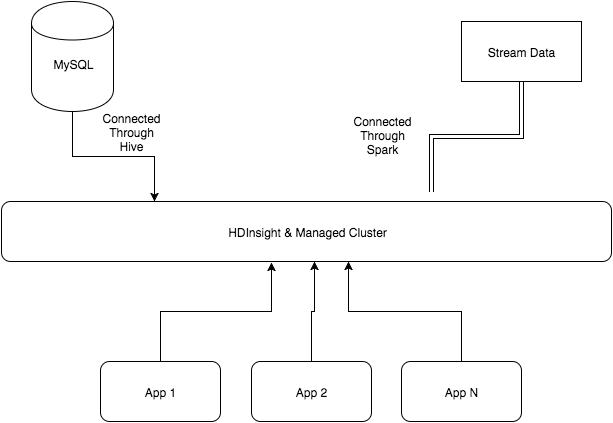
1. **Cortana Analytics Suite**
   1. **Cortana**: Digital personal assistant software. Capable of recognizing natural speech, recognizing and answering natural language questions, setting reminders, active listening, etc.
   2. **Power BI**: A suite of business analysis tools. Software includes a set of dashboards, reporting, data visualization, and real time analysis.
   3. **Azure Machine Learning**: Cloud based machine learning tools with drag and drop interface. Includes sets of algorithms for predictive analytics with custom R and Python support.
   4. **Azure HDInsight**: Provides scalable cluster management for software in the Hadoop ecosystem. Pricing models charge only for data and computational time used.
   5. **Azure Data Lake Analytics Service**: Platform built for developers and analysts to be a solution for all big data storage, processing, and analytics. Built on yarn for the cloud.
   6. **Azure Stream Analytics**: Made for performing real-time analytics for IOT solutions. Comes with dashboards and SQL-like language for quick development.
   7. **Azure Data Lake Store**: A HDFS for the cloud with no fixed limits on size. Made for massive throughput with high reliability and availability.
   8. **Azure SQL Data Warehouse**: Elastic data warehouse as a service capable of transacting SQL queries across relational and nonrelational data. Works seamlessly with HDInsight and other Data Lake services at the Petabyte scale.
   9. **Azure Data Factory**: Allows users to create, schedule and manage data pipelines. Comes with monitoring and visualization tools.
   10. **Azure Data Catalog**: Enterprise-wide metadata catalog. Allows easy access to data sets without having to move them.

**Azure Event Hubs**: Capable of connecting millions of devices across multiple platforms. Connections can then be streamed to other applications.

1. **Physical vs. Logical Layers**
   1. **Physical**

Combination on Managed VMs and Hadoop nodes on HDInsight. Applications on local or other machines in the cloud connect to HDInsight to access data.

* 1. **Logical**



1. **Administrative Tasks**
   1. **Requirements:** To be reviewed with Phillip
   2. **Git**: To be reviewed with Phillip