

The Lambda Architecture enables a continuous processing of real-time data. It is a painful process that gets the job done, but at a great cost. Splice Machine offers a simplified solution, Lambda-in-a-Box, that delivers the benefits of Lambda without the “enterprise duct tape” of other approaches.

Defining Traditional Lambda Architectures

Lambda Architectures are ubiquitous in machine learning and data science applications. The Lambda Architecture enables continuous processing of real-time data without the traditional ETL lag that plagues traditional operational (OLTP) and analytical (OLAP) implementations. In traditional architectures, OLTP databases are normalized for performance and then extensive ETL pipelines de-normalize this data, typically into star schemas on OLAP engines. This process usually takes at least a day. The Lambda Architecture circumvents this lag.

In its purest form, the Lambda Architecture sends data down two paths: a batch processing layer and a speed layer. The batch layer aggregates the raw data and trains models. The results of the batch layer are then transferred to a serving layer for the application to surface. The speed layer allows applications to access the most current data that “missed” the last batch processing window.

So how do companies implement this architecture? For the batch layer, they typically use a batch analytics processing engine on Hadoop, like MapReduce, Hive or Spark. For the serving layer, they use a NoSQL/Key-Value engine that performs well on batch imports and reads like Cassandra, HBase, Impala and Druid. For the speed layer, there typically is a queuing system like Kafka and a streaming system like Storm, Spark Streaming or even Flink.

5 Reasons Why You Need a Better Lambda Architecture

- 1. Simplify Operational Complexity**
 It is expensive to keep systems in sync and tuned for performance, and requires expertise in many technologies
- 2. Eliminate Need For Special Coding Skills**
 Developers have to write low-level code for basic operations and write batch programs to cleanse and filter individual records
- 3. Provide Access to Standard Tools**
 99% of data scientists know SQL and have used BI visualization tools, but traditional Lambda approaches throw away these powerful tools
- 4. Minimize Storage**
 Lambda typically requires at least duplicating the data in multiple engines
- 5. Support Integrated Applications**
 The modern application needs application logic co-mingled with analytical logic to be able to use raw data for business applications in the moment

Create Real-Time, Scalable Applications with a Simplified Lambda Architecture

Splice Machine offers a better solution to the complexity of Lambda Architectures. We call it Lambda-in-a-Box. With the new scale-out RDBMS systems, you can now get all the benefits of Lambda with a much simpler architecture.

For example, here's how a machine learning application can use Lambda-in-a-Box:

- **Batch File Ingestion** – Imports of raw data files are directly inserted into sharded tables in parallel with indexes that are atomically updated with the data for fast access
- **Real-time Stream Ingestion** – Stored procedures continuously ingest streams with standard SQL and auto-shards
- **Data Cleansing** – Use standard SQL, with constraints and triggers, to clean up small subsets of data as well as entire data sets efficiently, without big batch runs or file explosions
- **Feature Engineering and Extensive ETL** – Execute complex aggregations, joins, sorts, and groupings with efficient SQL that is automatically parallelized and optimized without writing code at the application level
- **Model Training** – Stored procedures execute analytics directly on the data, for example, using built-in functions like ResultSetToRDD that take SQL results and treat them as Spark RDDs or execute R and Python libraries directly on database result sets.
- **Application Logic** – ACID semantics enable the architecture to power concurrent CRUD applications without additional moving parts
- **Model Execution** – Stored procedures and user-defined functions wrap models
- **Reporting and Data Visualization** – Use Tableau, Domo, MicroStrategy and other ODBC/JDBC compliant tools turnkey

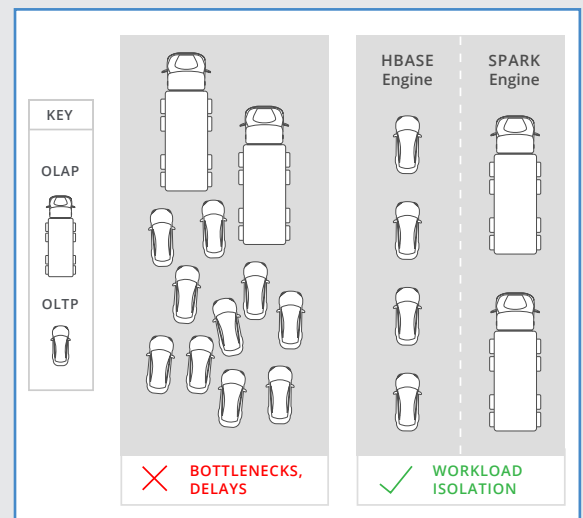
Why Choose Splice Machine for Your Lambda Architecture

Splice Machine is the open-source SQL RDBMS, powered by Apache Hadoop® and Apache Spark™.

The Splice Machine RDBMS provides:

- **ANSI SQL** – Splice Machine provides ANSI SQL-99 coverage, including full DDL and DML
- **ACID Transactions** – Splice Machine provides CRUD operations with fully ACID-compliant transactions, which scale both to high-frequency operational queries and very large-scale analytic workloads.
- **Secondary Indexing** – Splice Machine supports true secondary indices on data, in both unique and non-unique forms
- **Referential Integrity** – Referential integrity, such as Primary and Foreign key constraints, can be enforced without requiring any behavior from the underlying application
- **Joins** – Splice Machine supports inner, outer, cross and natural joins using join algorithms such as broadcast, merge, merge sort, batch nested loop, and nested loop joins

By centralizing on a relational Lambda Architecture on Splice Machine, teams can build machine learning applications very quickly, maintain them with standard operational personnel, and be able to tightly integrate machine learning into the application without extensive use of “Enterprise Duct Tape”.



Workload Isolation – The cost-based optimizer chooses a dataflow engine based on the estimation of the query plan – OLTP runs on HBase and OLAP runs on Spark