Apache Flume

Technology/Integration/Messaging/Data Collector

Description: A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It uses a simple extensible data model that allows for online analytic application. Flume is an Apache top-level project, Apache 2.0 license, written in Java.

Consequences:

★★ **Performance –** can handle high throughput with low latency, depends on channel configuration (memory, file etc.), number of sinks, threads etc. Supports horizontal scaling for performance (throughput) improvement.
★★ **Reliability** – uses a transactional approach to guarantee the reliable delivery of events (via a file channel). Usage of a memory channel improves performance, but can lead to message loss.
★★★ **Cost economy** - released as open source under the terms of the Apache License

Logstash

Technology/Integration/Messaging/Data Collector

Description: Logstash is a tool for managing events and logs. You can use it to collect logs, parse them, and store them in a centralized storage for later use (e.g. for searching). Logstash is a JRuby based application which requires the JVM to run. It is a part of the Elasticsearch family. The license is Apache 2.0.

Consequences:

★★ **Performance –** can handle high throughput with low latency (via leveraging multiple threads for filter workers, input/output workers etc., adding more instances)
★★ **Reliability –** depends on broker implementation, reported issues with losing messages
★★★ **Cost economy** - released as open source under the terms of the Apache License

Fluentd

Technology/Integration/Messaging/Data Collector

Description: Fluentd is an open source data collector (Apache 2.0 License), which unifies the data collection and consumption for a better use and understanding of data. Uses JSON as an internal format to unify all the collect, filter, buffer, and output mechanisms across multiple sources and destinations. Fluentd is written in a combination of C and Ruby.

Consequences:

★★ **Performance –** requires very little system resources. The vanilla instance runs on 30-40MB of memory and can process 13,000 events/second/core.
★★★ **Reliability** – supports memory- and file-based buffering to prevent inter-node data loss. The buffering logic is highly tunable and can be customized for various throughput/latency requirements.
★★★ **Cost economy** - released as open source under the terms of the Apache License

RabbitMQ

Technology/Integration/Messaging/Distributed Message Broker

Description: RabbitMQ is open source message broker software (sometimes called message-oriented middleware) that implements the Advanced Message Queuing Protocol (AMQP).

Consequences:

★★ **Performance –** good response rate and quite high memory consumption, scales with clustering well

★★ **Reliability** – implemented distributed from scratch (Erlang), built-in clustering, can persist and replicate configuration and messages, allows acknowledgements. Can exhaust RAM.

★★★ **Cost economy** - released as open source under the terms of the Mozilla Public License

Apache Kafka

Technology/Integration/Messaging/Distributed Message Broker

Description: Apache Kafka is an open-source message broker project developed by the Apache Software Foundation written in Scala. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds.

Consequences:

★★★ **Performance –** very fast response rate, scales with clustering

★★★ **Reliability** – durable and fault tolerant by design and implementation language (Scala), persists and replicates messages

★★★ **Cost economy** - released as open source under the terms of the Apache License

Amazon SQS

Technology/Integration/Messaging/Distributed Message Broker

Description: Amazon Simple Queue Service (SQS) is a fast, reliable, scalable, fully managed message queuing service. SQS makes it simple and cost-effective to decouple the components of a cloud application.

Consequences:

★★½ **Performance –** serves up to 35K msg/sec, auto-scales by Amazon

★★★ **Reliability** – designed for high availability, works as a service, very reliable, durable messages

★★ **Cost economy** - $0.50 per 1 million requests per month

Apache ActiveMQ

Technology/Integration/Messaging/Distributed Message Broker

Description: Apache ActiveMQ ™ is an open source messaging and Integration Patterns server. Apache ActiveMQ is fast, supports many Cross Language Clients and Protocols, it comes with easy to use Enterprise Integration Patterns and many advanced features while fully supporting JMS 1.1 and J2EE 1.4.

Consequences:

★★ **Performance –** good response rate on par with RabbitMQ but slower than Kafka

★★ **Reliability** – high availability is supported, persistent messaging supported, some issues reported with clustering and occasional message loss

★★★ **Cost economy** - released as open source under the terms of the Apache License

Apache Cassandra

Technology/Data Storage/NoSQL Database/Column-Family

Description: The Apache Cassandra is an open source distributed database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple datacenters, with asynchronous masterless replication allowing low latency operations for all clients.

Consequences:

★★★ **Performance** – Cassandra is 30%-100% faster (avg) than HBase for both reads and writes due to efficient memory/caching, SSD support, online snapshots, locally-managed storage, effective compaction, etc. It is a winner of most performance benchmarks in its class.
★★★ **Reliability** – one of the most reliable and mature NoSQL databases
★★★ **Real-time analysis** – fast access to data makes it a good solution for real-time analysis
★★★ **Cost economy** - released as open source under the terms of the Apache License

Apache HBase

Technology/Data Storage/NoSQL Database/Column-Family

Description: HBase is a NoSQL databases which uses HDFS and Hadoop as a storage engine, and integrates well with most Hadoop products. Typical use case is random, real-time read/write access to your Big Data, scaling up data to billions of rows, and millions of columns. Apache HBase is an open-source, distributed, versioned, non-relational database modeled after Google's BigTable.

Consequences:

★★½ **Performance** – is still fast compared to other databases, but slower than Cassandra due to underlying HDFS, ineffective compaction, etc.
★★ **Reliability** – very slow crash recovery, compactions that disrupt operations, unreliable splitting, and very small write-ahead logs makes it less reliable than Cassandra despite reliable HDFS storage
★★★ **Real-time analysis** – one of the fastest random access to data in NoSQL world makes it a good solution for real-time analysis
★★★ **Cost economy** - released as open source under the terms of the Apache License

MongoDB

Technology/Data Storage/NoSQL Database/Document-Oriented

Description: MongoDB (from "hu**mongo**us") is an open-source document database, and the leading NoSQL database. Written in C++, MongoDB features: JSON-style documents with dynamic schemas offer simplicity and power, indexes on any attribute, mirrors across LANs and WANs for scale, scales horizontally without compromising functionality, flexible aggregation and data processing, etc.

Consequences:

★★ **Performance** – not as fast as simplest key-value storages, but features like auto-sharding, full index support, map-reduce makes it fast enough. Written in C++.
★★ **Reliability** – durability is a known problem (being fixed though), issues with repairing databases, requires replication setup to implement reliability
★★★ **Real-time analysis** – one of the most common use-cases, supports schema design, indexing and sharding for real time analytics workloads
★★★ **Cost economy** - released as open source under the terms of the GNU AGPL license, commercial licenses are also available

Apache CouchDB

Technology/Data Storage/NoSQL Database/Document-Oriented

Description: CouchDB is a database that embraces the web by storing data with JSON documents; allowing accessing data via HTTP; indexing, combining, and transforming your documents with JavaScript. CouchDB works well with modern web and mobile apps, supports incremental replication and master-master setups with automatic conflict detection.

Consequences:

★½ **Performance** – fast for direct ID lookups and map-reduce jobs, but that's it. Users reported performance issues.
★ **Reliability** – serious problems with reliability and availability were reported by users despite functionality like replication and automatic conflict resolution. Not yet suitable for highly-available or heavy-loaded solutions.
★★★ **Real-time analysis** – fast ID lookups and fast aggregation calculation using map-reduce
★★★ **Cost economy** - released as open source under the terms of the Apache License

Impala

Technology/Analytics/Search & Query/Interactive Query Engine

Description: Cloudera's open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop.

Consequences:

★★★ **Performance –** considered as one of fastest technologies at the moment, significantly faster than Hive

★★ **Processing capabilities** – supports the SQL-92 standard, but overall features are limited compared to HiveQL

★★ **Reliability** – designed for short queries; queries must be restarted if a node fails

★★★ **Cost economy** - released as open source under the terms of the Apache License

Spark SQL

Technology/Analytics/Search & Query/Interactive Query Engine

Description: Based on Spark – an in-memory distributed computing engine (alternative to Hadoop MapReduce), Spark SQL allows running SQL and HiveQL queries over large datasets. Spark SQL is an ancestor of Shark.

Consequences:

★★★ **Performance –** considered as one of fastest technologies at the moment, significantly faster than Hive

★★ **Processing capabilities** – based on SQL-like query language supporting most of HiveQL features including UDFs and SerDes

★★★ **Reliability** – supports long-running queries and mid-query fault recovery

★★★ **Cost economy** - released as open source under the terms of the Apache License

Apache Hive

Technology/Analytics/Search & Query/Interactive Query Engine

Description: The Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage.

Consequences:

★½ **Performance –** even with Stinger initiative Hive is still slow compared to other alternatives such as Impala or Spark SQL

★★★ **Processing capabilities** – are based on HiveQL, a subset of SQL-92 which offers extensions such as non-scalar data types, XML/JSON functions, UDFs, custom SerDes and other features

★★★ **Reliability** – supports long-running queries and mid-query fault recovery

★★★ **Cost economy** - released as open source under the terms of the Apache License

Splunk (Indexer)

Technology/Analytics/Search & Query/Distributed Search Engine

Description: Splunk Indexer is a component responsible for indexing and correlation of real-time data in a searchable repository. Along with other Splunk components it creates a larger solution which provides graphs, reports, alerts, dashboards and visualizations.

Consequences:

★★½ **Ad-hoc analysis** – provides REST-like API and CLI with proprietary SPL language, supports faceted search, abnormalities finding, SQL-like joins, stats functions and functions that support chart visualization

★★½ **Real-time analysis** – enables near real time indexing with about a second latency

★★★ **Reliability** – implements sharding, load balancing, automatic failover and recovery

★ **Cost economy** - the freeware version is limited to 500 MB of data a day, $1,800 for a 1 GB/day annual term license or $4,500 for a 1 GB/day perpetual

Elasticsearch

Technology/Analytics/Search & Query/Distributed Search Engine

Description: An open source search and analytics engine based on Lucene. It provides distributed, multitenant-capable full-text search capabilities with a RESTful web interface and schema-free JSON documents. Along with Logstash and Kibana creates ELK stack to collect, index and visualize data.

Consequences:

★★ **Ad-hoc analysis** – provides JSON-based query language Query DSL, supports aggregations and filtering, allows joins through *has_child* queries

★★½ **Real-time analysis** – enables near real time indexing with about a second latency

★★★ **Reliability** – implements sharding, load balancing, automatic failover and recovery

★★★ **Cost economy** - released as open source under the terms of the Apache License

Apache Solr

Technology/Analytics/Search & Query/Distributed Search Engine

Description: An open source enterprise search platform from the Apache Lucene project. Providing distributed search and index replication, Solr is highly scalable and adopted by a number of Big Data vendors such as Cloudera and Amazon Web Services.

Consequences:

★★ **Ad-hoc analysis** – provides REST-like API and extends Lucene query language, supports faceted search and filtering including pivot facets

★★½ **Real-time analysis** – enables near real time indexing with about a second latency

★★★ **Reliability** – implements sharding, load balancing, automatic failover and recovery

★★★ **Cost economy** - released as open source under the terms of the Apache License