INTRODUCTION TO CASSANDRA

This eBook provides a high level overview of Cassandra and describes some of its key strengths and applications.
WHAT IS CASSANDRA?

Apache Cassandra is a high performance, open source database designed to handle very large amounts of data. It’s an economical solution that provides a distributed architecture for continuous availability and no single point of failure, and offers superior reliability and scalability.

Cassandra is a NoSQL database management system (DBMS), or a database that doesn’t require that the data be in the tabular format required of traditional relational database management systems (RDBMS) and doesn’t need to use standard SQL for data storage and retrieval. NoSQL databases offer a high level of flexibility, so IT departments are always ready to deal with new types of data as they emerge, regardless of volume or structure. This flexible structure makes them well suited for web and mobile applications, as well as the for Internet of Things (IoT) applications. Other NoSQL databases include HBase, MongoDB, and Couchbase.

History of Cassandra

Cassandra was developed at Facebook to support its Inbox search capabilities. It was officially introduced as an open source technology in 2008 as Apache Cassandra, and a couple of years later, redistributed by DataStax as an enterprise offering. For companies that need more advanced capabilities and full product support, Datastax Enterprise extends Cassandra with built-in analytics, search, and security capabilities, along with formal training and product support.
How it’s used

Cassandra enables data storage across as many commodity servers as required in order to distribute workload and ensure performance levels that could not generally be guaranteed with a traditional RDBMS. This can greatly reduce the hardware costs associated with large server purchases and makes it easier to accommodate high-volume data writes. This distributed architecture ensures that there is no single point of failure as data is replicated across the Cassandra nodes, or instances. In other words, if one of the servers experiences an issue, the database can automatically redistribute the load so that neither access nor performance are impacted. This makes Cassandra an ideal solution for scale-out strategies as data volumes grow.

For companies that are making use of cloud computing, Cassandra is an ideal solution. It integrates easily with cloud computing platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Compute Engine.

While Cassandra is the superior platform for a high-volume data writes, for use as a solution for data retrieval and analytics you can pair it with a distributed storage and processing platform such as Spark or Hadoop. The Spark platform is particularly well suited for fast data retrieval because of its in-memory processing. Hadoop is recommended when you need to handle very large data volumes.

Who’s using Cassandra

As Netflix grew in popularity, they needed to eliminate downtimes in their SQL database in order to ensure a positive and consistent user experience. Database issues and frequent schema changes were causing some services to become unavailable for up to 10 minutes at a time, which was not acceptable for the service they aimed to provide to their customers. Netflix now uses Cassandra running on Amazon Web Services (AWS) to store customer and media information such as ratings. This implementation has removed the possibility of any single point of failure, which ensures that they have a highly available service at all times.
eBay deployed Cassandra to track buying history and user activity. Using this information, they can tailor the eBay user experience for each individual customer. With Cassandra, their data store is always available, and the personalized shopping experience is never compromised even with the extremely high volumes of data that the eBay system must process each day.

At Spotify, Cassandra manages information on customer playlists and product catalogs, which are being created at very high volumes as a growing number of users access the application. With this data, Spotify can personalize every user’s experience and automatically present him/her with music playlists that are tailored to individual preferences.

This represents just a small sample of the companies that use Cassandra and DataStax to support some critical aspect of their business. Across these examples, one commonality is the requirement for reliable performance in the face of massive amounts of data and the ability to scale as these demands grow.

**Practical Applications**

NoSQL databases can be loosely classified into four main categories: Column, Document, Key Value and Graph. Cassandra is unique among NoSQL solutions because it’s a Columnar/Tabular database. This means data is stored in flexible columns and tables instead of rows, as with a RDBMS. With this type of database, storage is much less structured than a RDBMS, but it enables a high level of flexibility and exceptional performance.

When considering whether Cassandra may be a good fit for your project, it’s beneficial to understand that there are certain scenarios in which Cassandra is especially well suited. You may be considering a Cassandra deployment when you have:

- Data volumes that are measured in terabytes or petabytes
- A need to ingest or write the data at very high speed
- Data that needs to be geographically distributed
• A requirement for high fault tolerance
• A need to reduce database costs
• At least some data that’s unstructured or semi-structured (although not schemaless)
• A solution that must easily scale as data volumes grow
• A need for high performance

Cassandra is best suited when fast write performance or storage is required — for example, when collecting metrics or shopping cart data. This data can be analyzed at a later time, but analytics is not the primary application of Cassandra.

Challenges and Considerations

When an IT department decides to use Cassandra for a particular business application, the initial configuration of the database is extremely important to the overall success of the project. It is important that the data model is designed and implemented appropriately from the outset to ensure maximum scalability and performance. Also, it’s critical to monitor the environment once it has been implemented. Because the database will continue to function regardless of node failures, it’s necessary to detect and resolve failures as quickly as possible in order to keep the system at peak health.

Since Cassandra does not rely on the data structure and query language of a traditional RDBMS, it requires a set of unique technical skills to ensure peak efficiency and stability. These skills include in-depth experience in data modelling and administration, and an solid understanding of how to monitor the system. And since it is a relatively new technology, companies who have decided to use it may find it difficult to attract and maintain administrators with the expertise required to manage this essential IT resource.
Achieving Success with Cassandra

In order to best take advantage of the many benefits of deploying Cassandra and derive the most value in the shortest possible timeframe, it’s a good idea to partner with someone who can provide highly knowledgeable and experienced Cassandra database experts. You need professionals who can step in and work closely with your team at any stage of your project. This approach allows you to address critical skills and capacity gaps within the team and get your Cassandra instance up and running quickly, and performing optimally as you move forward. Here are some key milestones of your implementation for which you may consider engaging specialized assistance:

- NoSQL education and technology selection
- Initial assessments against an existing Cassandra environment if it exists
- Cassandra installation and configuration
- Cassandra database modelling and development
- Operational support for your Cassandra environment

If you’re thinking of implementing Cassandra, watch our webinar, Getting Started With Cassandra, which covers key topics for starting out, such as when you should use Cassandra, potential challenges, real world Cassandra applications and benefits, and more.

If you would like more information, visit www.pythian.com or contact us by email at info@pythian.com