EXECUTIVE BRIEF

OPEN SOURCE BIG DATA PROJECTS - EMERGENCE OF THE CONVERGED DATA PLATFORM

THIS EXECUTIVE BRIEF IS A SUMMARY OF THE WEBINAR:

The Big Data Blender: Converging Hadoop, Spark, Streaming, and More

CLICK TO ACCESS THE WEBINAR

INTRODUCTION

- As of Jan 2016, 451 Research identified over 275 vendors and products in the data platform and analytics landscape. Growth is expected, as is convergence of data platforms.
- Convergence or the 'blending' of data management platforms is being driven by
  1. Different data types - the need to adapt some data stored in different types of databases
  2. Operational efficiencies - reducing maintenance by converging multiple data silos into one
  3. Demands of variable workloads – some data stores are better at addressing certain types of applications and their workloads
- Open source technology is increasingly used in the growing complexity of Big Data environment projects. Although able to handle scale and complexity of the new modern data types, a converged data platform allows it all to be delivered on a unified platform, providing centralized management, security, high availability, fault tolerance and disaster recovery.

DEFINITION

- A converged data platform is the convergence or blending of two or more processes, frameworks or technology - coming together in a unified whole.

IMPORTANT DATA

- By 2019 it is anticipated that the value created in the Hadoop, Event/Stream processing, New SQL and NoSQL market will be about $10b. Open source Hadoop will account for about $3.6b and NoSQL some $4.7b. 451 Research expect continued growth in this market.

KEY POINTS

- NoSQLs emergence, growth and appeal have been attributed to its different data stores for different workloads (polyglot persistence). However, limitations of operational complexity and inflexibility from multiple databases driving applications have been addressed by the growth of multi-model databases - which support a combination of various NoSQL data models. NoSQL's future is predicted to rest with this approach.
- With increasing complexity in data workloads, a number of actions and data sets get joined at many management points that require different consoles, different hardware utilization, different demands for security and different fault tolerance and disaster recovery. These concepts drive the creation of a converged data platform.
- In the past operational and analytical data workloads were more or less separated. Today they often are not and this demands data convergence.

KEY POINTS (cont.)

- Convergence is also being driven by the idea that some in-memory databases can also serve as memory caches. Thus platforms are being developed that can run a database with a cache or simply as a cache solution.
- The expansion of streaming technology has opened up yet another area for convergence where diverse applications are being developed to serve expanding demands for analysis of real-time data streams.
- Total Data Warehouse is a term used by 451 Research to describe all the components that make data platform convergence possible.

BUSINESS BENEFITS

- Converged data platforms are transforming businesses, demonstrating significant cost savings and helping drive additional revenues. Six cases studies are cited by MapR in their recent webinar with 451 Research.
- The key benefits of a converged data platform:
  1. It's in real time with reduced latency - improves responsiveness of applications
  2. Improved reliability drives greater business value and reduced costs
- Converging data technologies in one place creates between 30% to 50% reduction in the overall total cost of ownership. Savings are evident on the type of hardware utilized, the data center, operating costs, and reduced costs of data movement.
- In a survey of MapR customers, it was reported that 98% were running more than one application on a single cluster and 18% were running over 50 applications.

TRENDS

- Different data types (structured, unstructured), operational efficiencies (converging multiple silos of data) and variable workloads (dependant on applications) are all driving convergence of data storage platforms.
- NoSQL and SQL vendors are starting to adopt certain traits of each other. For example, IBM and Oracle are adding JSON capabilities to their databases, searchable by SQL. NoSQL vendors are adding SQL querying capabilities. This trend illustrates NoSQL and SQL convergence.
- Operational and analytical data workloads are starting to converge. Until recently, separate databases have been used for each. Emerging databases take advantage of in-memory and advanced processing to deliver combined operational and analytical processing.
- Several NoSQL players, among them MapR, have taken up the multi-model approach and it is anticipated that this is the future direction of this sector.
- Hadoop based convergence is taking place in
  1. Cache (Apache Geode, Apache Ignite)
  2. Operational databases (NoSQL) - (MapR-DB, Apache HBase, Splice Machine, Apache Trafodian)
  3. Analytic databases, made possible by connectors, SQL-on-Hadoop, Federated query - (e.g. Pivotal, Teradata and IBM)

TRENDS (cont.)

- Convergence is also occurring with cache and grid databases where databases can be run with a cache or used as a pure cache solution.
- Customers' growing need for more frequent analysis of real-time data streams is pushing data stream processing into the mainstream. The blending of streaming technologies (e.g. Storm, Spark, Kafka and MapR Streams) with Hadoop illustrate the convergence in this space.
- JSON is an extremely popular format among application developers and it is emerging as a de facto standard for storing data, particularly around the growth of sensors for IoT use cases.
- OJAI (Open JSON Application Interface) is seen by MapR as an emerging standard for managing JSON data access across the entire data platform, whether that is the database itself, streams or file systems.

MapR'S OFFERING

(Key points on company offering made by MapR in the webinar)

- MapR provides the industry's only converged data platform that integrates the power of Hadoop and Spark with global event streaming, real-time database capabilities, and enterprise storage, enabling customers to harness the enormous power of their data. Organizations with the most demanding production needs, including sub-second response for fraud prevention, secure and highly available data-driven insights for better healthcare, petabyte analysis for threat detection, and integrated operational and analytic processing for improved customer experiences, run on MapR.
- MapR's approach is to be as open as possible in supporting as many of the API's as possible. One of the very popular features of MapR's converged data platform is the seamless movement of data and files through drag and drop without having to write special code to batch-load data into and out of a Hadoop environment.
- MapR's converged data platform brings together open source engines and tools (e.g, Hadoop, Spark, Apache Drill) and can also support commercial engines and applications (e.g. Vertica, SAP, MySQL).
- MapR's converged data platform handles multi-model databases. JSON is a recently added standard native format on which applications can also be built. MapR is completely integrated with Hadoop and can operate across datacenters in a multi-master type of environment and setup.
- In the area of security management across a big data environment, MapR has pushed down access control to the granular level. This access control can be spread across different processing engines and all types of structured data, files, tables and streams.
- The MapR system is a multi-tenant environment supporting many different applications on a single cluster.
- In realtime applications, MapR's data platform allows a single system to operate simultaneously with data-at-rest and data-at-motion reducing latency.
- MapR's converged data permits an optimized table setup and streaming working directly with storage hardware resources. It operates natively against the hardware allowing fast, efficient, direct I/O on that system, thereby improving performance.

MapR'S OFFERING (cont.)

- MapR Streams is a global streaming service that extends the capacity of Sparks, Storm and other streaming processing engines. It pushes billions of messages per second in a reliable way thus allowing information consumption to be virtually real time.
- The fulfillment of the promise of docker containers and the ability to move workloads on the fly to different data center nodes is handled seamlessly in MapR.

CASE STUDY
(Cited by MapR in the webinar)

- The largest healthcare provider in the United States (United Health Group) run what they call a 'big data as-a-service' platform using MapR's converged data platform and have saved hundreds of millions of dollars in improving efficiency and reducing waste in how they manage and process insurance claims.

TAKEAWAYS

- The business value of converging different data workloads (e.g, batch, interactive) and streaming them across the business, manifests itself in significantly improved customer experiences and value.
- There is considerable cost reduction with data convergence where data sprawl and data duplication are controlled. Administrative costs associated with overall enterprise data architecture are reduced.
- The converged data platform not only simplifies app development but also the way apps are run in a data center.
- With new customer demands and improving technology capabilities, all roads lead to a converged data platform.
- MapR has created a simple process of getting started with its converged data platform. Free on-demand training is available for Hadoop, Spark and SQL engines. Quick start solutions using blueprints and templates are available so that MapR clients can be operating with a 6 node environment in 4 to 6 weeks.
- Questions to ask your Big Data vendor:
  1. Is my data highly available? Can we plug in existing enterprise systems?
  2. Can we properly identify users?
  3. Is multi-tenancy supported?
  4. Is my data correct and supported?
  5. Can we authorise access to data?
  6. Are apps supported across geographies and data centers?
  7. Is my data governed?
  8. Is there a proper paper trail?

CLICK TO ACCESS THE WEBINAR

ABOUT MapR

MapR provides the industry's only converged data platform that integrates the power of Hadoop and Spark with global event streaming, real-time database capabilities, and enterprise storage, enabling customers to harness the enormous power of their data. Organizations with the most demanding production needs, including sub-second response for fraud prevention, secure and highly available data-driven insights for better healthcare, petabyte analysis for threat detection, and integrated operational and analytic processing for improved customer experiences, run on MapR.

ABOUT 451 RESEARCH

With a core focus on technology innovation and market disruption, 451 Research provides essential insight for leaders of the digital economy. More than 100 analysts and consultants deliver that insight via syndicated research, advisory services and live events to over 1,000 client organizations in North America, Europe and around the world. 451 Research and its customers benefit from the combined assets and talent of The 451 Group and its two divisions: 451 Research and Uptime Institute.