Security for Big Data

A Dell Big Data White Paper

By Joey Jablonski

The power to do more

Security for Big Data

As more and more companies deploy big data technologies, including Apache Hadoop[1], Cassandra[2] and other related technologies, security is becoming more critical. As big data technologies become more mainstream, it is critical that they are deployed with the same safeguards, auditing and protection capabilities inherent in existing IT platforms such as BI tools, RDBMS platforms and data storage platforms. Because of the relative new-ness of big data platforms, the security community is working rapidly to create the necessary capabilities for seamless integration into existing security frameworks. Gaps do still exist, but are rapidly being closed. These gaps demand focus by organizations as they manage risk related to security and access of corporate data assets.

Planning security for a big data solution is similar to that of other data-centric platforms deployed by many IT departments. The striking difference is two-fold:

1) ***Maturity of the technology*** – Many Big data technologies are less then 10 years old, while existing relational database management systems have been around for 20+ years. This lifecycle leads to fewer features and less flexibility in how big data platforms are deployed and integrated with existing systems.
2) ***Pan-organizational data in a single platform*** – Big data platforms, like Hadoop, commonly become an integration point across other existing systems. This leads Hadoop to store data sets together that had not previously been combined. This combined data can introduce new risks around who accesses it, when, and the level of risk associated with data that is more detailed than other data silos.

Big Data Security Challenges

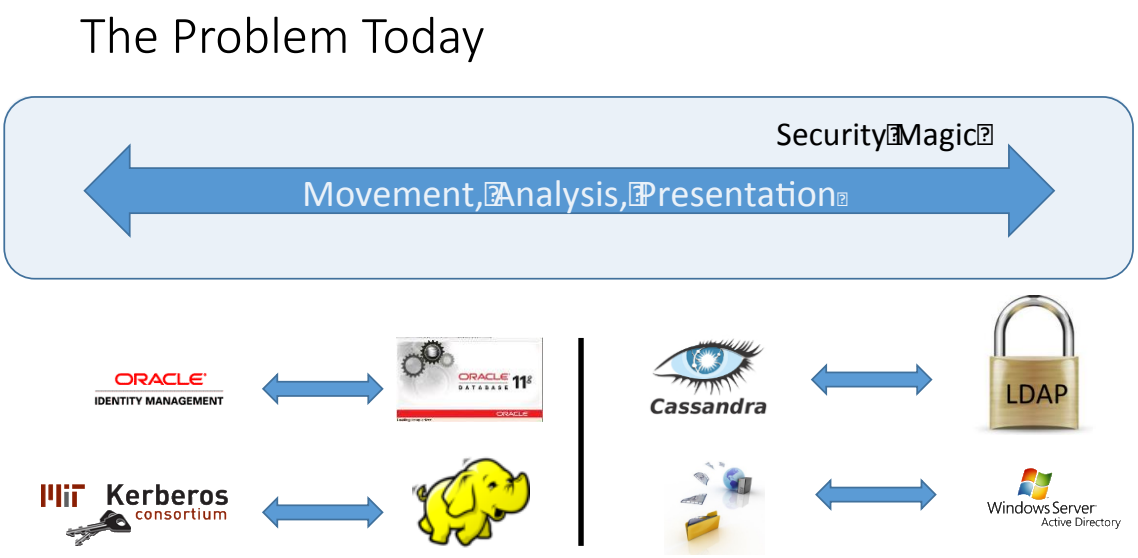

Figure 1 - Todays Big Data Security Challenge

As shown in Figure 1, many big data projects today are plagued by struggles with integration of the modern technology platforms into existing mechanisms for access control, authentication and authorization. Today's big data technologies, including platforms like Hadoop and Cassandra, have the ability to be integrated with existing identity and access control systems like Lightweight Directory Access Protocol (LDAP), Kerberos and Active Directory. All newly deployed platforms should leverage existing implementations for access control to ensure uniformity across silos that store and house data.

Often times new big data deployments demand that organizations make a tradeoff from their existing security policies, and new technologies that have not matured to the point of being able to implement access controls as a company has defined or prefers. It is important that each organization weighs their own risk profile in this situation and determines if they are best fit to deploy this new technology for organizational benefit, or if the risk of data compromise is too great that deployment of the technology should wait until it has matured further.

Big Data Design Considerations

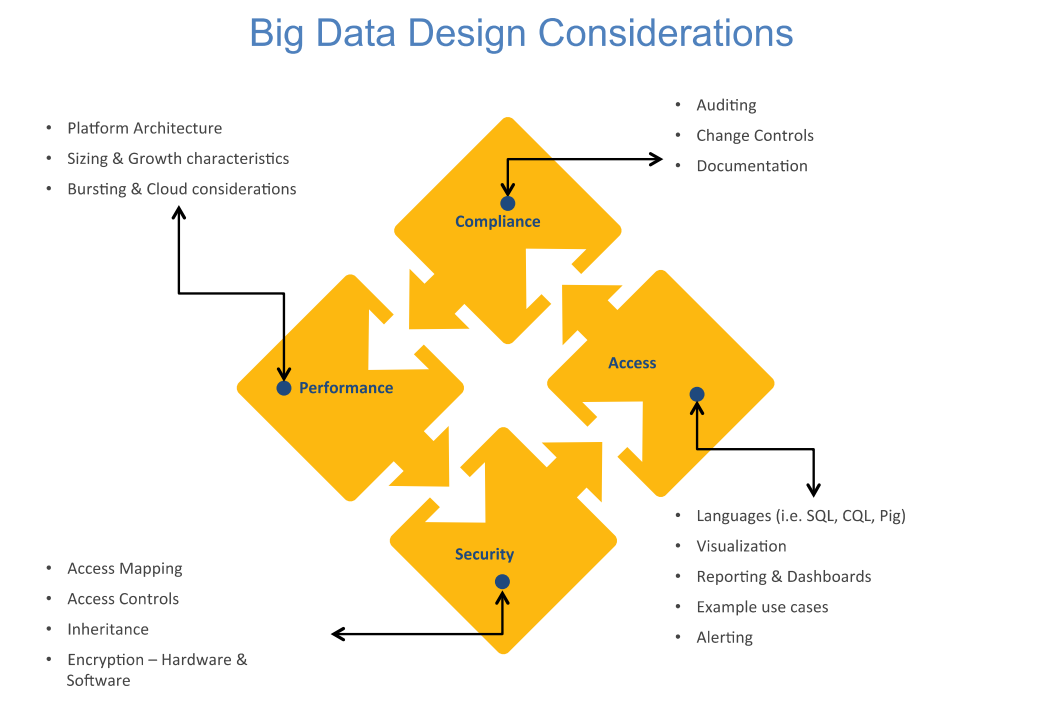

Figure 2 - Big Data Design Considerations

Figure 2 outlines the key considerations for designing any big data environment. There are four key areas of design consideration, all equally weighted in a final solution deployment:

- ***Performance*** – Any system for analyzing data and enabling decisions must be designed with performance in mind, this ensures that all users get access to data in the time needed to make effective decisions that positively affect the direction of the organization.
- ***Compliance*** – Compliance is a key component of all big data system designs. Compliance is the assurance that defined policies can be reported on and adverse effects acted upon.
- ***Security*** – Security, and the key focus of this white paper, includes the controls, safeguards and supporting technologies to control access to data and restrict it to authorized users, applications and processes.
- ***Access*** – Access is the alignment of user preferences with supplied technology for the use and presentation of the data stored in the big data solutions and platforms. Access is about ensuring the right user interfaces are available for consumption of the data.

Change of Paradigms in Big Data

Today's big data platforms, like Hadoop, have the ability to integrate a multitude of data from different sources and different types. Particularly in Hadoop, it is common to have a mix of structured and unstructured data, as well as primary copies of information supplemented by secondary copies from other systems. This mix introduces specific challenges that must be addressed in system design and implementation:

- ***Analyze versus view*** – Many traditional data systems were responsible for responding to user requests to provide access to data, very little analysis or transformation occurred. Today, many big data platforms are asked to both analyze data and present information. Access controls can be different for these activities, it may be that a user has the appropriate access to analyze a data set of de-identified data, but does not have the ability to view the individual records of that dataset. Big data implementations should account for this difference and ensure users can properly analyze data, while not exposing details that are unnecessary for carrying out the business decisions from that information. One method to enable consistent views of data is through "tokenization." Tokenization in data security is the process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security. Tools like Accumulo [3] enable all data to be tagged, as well as users and access carried throughout the process of encryption, decryption, analysis and presentation.
- ***Validation of Input, Validation of Results*** – Many big data platforms pull data from a large variety of sources. These sources can contain varied qualities of data that is then used for analysis and processing. All systems should have checkpoints in all workflows to ensure ingested data is of high quality and valid data. Invalid data should be discarded including data from unvalidated sources so that results are not adversely affected by users that do not have clear visibility into data sources.
- ***Data with a date component*** – Much of today's data created has a date component and often that date component can affect access levels. Financial information is the most common, the current quarter of financials is often much more sensitive because it cannot be shared publicly except under

specific conditions. Big data platforms should always factor in creation dates as part of tracking data lineage and access.

- ***Data + Data*** - Traditional data systems like databases were very effective at isolating data types through the use of tables and instances. Modern big data platforms do not have this separation, leading to instances where two pieces of data, when combined, require a higher level of protection than the individual components. This Data + Data problem, leads to organizations having to be conscious of how people access and combine data, and leads IT departments to think about how data will be exposed to end users and what access controls are required to protect data at the higher levels required.
- ***Documents with mixed data levels*** – Many documents today stored in big data environments continue multiple different levels of information that should be protected at different levels. Because these documents are unstructured, it is challenging to identify which data is of which classification level. The most common methods for protecting this information is to protect documents at the highest level of data they contain, but also creating replica copies of the files that are manually scrubbed of confidential information to allow wider access to a subset of the file.
- ***System Integration*** – As more and more systems become integrated with data flowing among them, it is critical that data models and access controls be consistent across systems. This ensures uniformity in data access for users that access similar data from different systems.
- ***Single View across access types*** – Tools like Hadoop provide many different methods for accessing data, some of these tools include Apache Hive[4], Apache HBase[5] and Apache Pig[6]. Each of these tools has the ability to create and manage independent sets of access controls. Tools like Apache Sentry[7] enable a uniform set of access controls across tools; this ensures lower risk data access, easier management of data access and uniform auditing of data access by users.

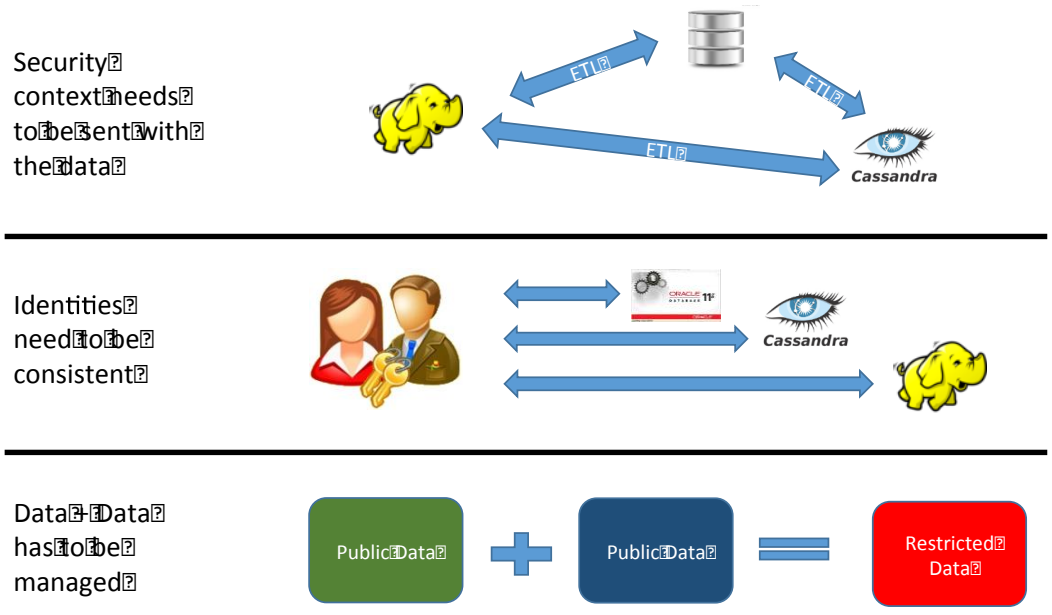

Figure 3 - Big Data Security Best Practices

As shown in Figure 3, Dell recommends three best practices for designing security implementations for big data solutions:

- ***Security Context*** – It is recommended that any time data is transmitted between systems that the context of that data is carried as well, including access policies, data lineage and aging policies.
- ***Identities*** – Identities should be managed across an organization with a centralized identify database used for verifying access to data in big data platforms. This ensures uniform access and compliance reporting across data silos in an organization.
- ***Data + Data*** – As part of the data access policies, a record of data sets and the implications of their integration should be tracked and updated on a regular basis. This record can be used to modify access controls on the fly and report on possible risks created by integrating data in new platforms and systems.

In addition to the above, it is critical that organizations standardize across all departments and teams on how to tag data. A standard tagging mechanism ensures that access controls and auditing can be applied uniformly. Users are the best source of this tagging as they can provide feedback real-time on data they create or access that is untagged.

Compliance is an important part of the security of data. Strong tools should be deployed as part of any big data solution to ensure that all data access and use can be reported on, and alerts generated for inappropriate data access. As data sets become more complex and more desperate data sets are integrated, ensuring compliance will become more difficult, but can be managed if data is integrated in steps, rather then all at once.

Security is a key component of all big data projects. All solution designs should encompass Performance, Access, Compliance and Security. Security should be defined at all levels of the system implementation and account for both at-rest and in-flight data. Big data systems introduce new challenges for security that must be accounted for including data + data policies and the handling of documents that may contain multiple levels of security. All big data projects should start small, with low risk data being the focal point; this enables organizations to get comfortable with new technologies as well as how best to ensure it can be implemented to conform to corporate security policies.

References

[1] Apache Hadoop - http://hadoop.apache.org/

[2] Cassandra - http://cassandra.apache.org/

[3] Accumulo - https://accumulo.apache.org/

[4] Apache Hive - https://hive.apache.org/

[5] Apache hbase - http://hbase.apache.org/

[6] Apache Pig - http://pig.apache.org/

[7] Apache Sentry - http://sentry.incubator.apache.org/

To learn more

To learn more about Dell big data solutions, contact your Dell representative or visit:
www.Dell.com/bigdata
www.DellBigData.com