

COMPUTERWORLD

IT must prepare for Hadoop security issues

Aggregating data from multiple sources can cause access control and data entitlement problems, analysts say

By Jaikumar Vijayan

November 9, 2011

Computerworld - NEW YORK -- Corporate IT executives need to pay attention to numerous potential security issues before using Hadoop to aggregate data from multiple, disparate sources, analysts and IT executives said at the [Hadoop World conference](#) here this week.

Open source [Hadoop technology](#) lets companies collect, aggregate, share and analyze huge volumes of structured and unstructured data from enterprise data stores as well as from weblogs, online transactions and social media interactions.

A growing number of firms are using Hadoop and related technologies such as Hive, Pig and Hbase to analysis analyze data in ways that cannot easily or affordably be done using traditional relational database technologies.

JPMorgan Chase, for instance, is using Hadoop to improve fraud detection, IT risk management, and self service applications. The financial services firm is also using the technology to enable a far more comprehensive view of its customers than was possible previously, executives said.

Meanwhile, Ebay is using Hadoop technology and the Hbase open source database to build a new search engine for its auction site.

Analysts said that IT operations using Hadoop technology for such applications must be aware of potential security problems.

Using the technology to aggregate and store data from multiple sources can create a whole slew of problems related to access control and management as well as data entitlement and ownership, said Larry Feinsmith, managing director of JPMorgan Chase's IT operation.

Hadoop environments can include data of mixed classifications and security sensitivities, said Richard Clayton, a software engineer with Berico Technologies, an IT services contractor for federal agencies.

The challenge for enterprises is to ensure that they implement appropriate security controls for enforcing role-based access to the data, he added.

Aggregating data into one environment also increases the risk of data theft and accidental disclosure, Clayton said.

And importantly, he noted, applications analyzing merged data in a Hadoop environment can result in the creation of new datasets that may also need to be protected.

Many government agencies are putting Hadoop-stored data into separate 'enclaves' to ensure that it can be accessed by those with clearance to view it, Clayton said.

Several agencies won't put sensitive data into Hadoop databases because of data access concerns, Clayton said. Several agencies are simply building firewalls to protect Hadoop environments, he added.

For many Hadoop users, the most effective security approach is to encrypt data at the individual record level, while it is in transit or being stored in a Hadoop environment, Clayton said.

In general, Clayton advises enterprises is to be cautious in their use of Hadoop technologies. He notes that built-in Hadoop Distributed File System (HDFS) security features such as Access Control Lists and Kerberos used alone are not adequate for enterprise needs.

"Security and access control are part of the reason why Hadoop is not ready to replace relational databases" in the enterprise, said David Menninger, an analyst with Ventana Research.

Sid Probst, the chief technology officer at Attivio, a vendor of unified access management technologies in big data environments, added that "Hadoop is a great technology but there is a whole series of enterprise readiness problems."