CITO Research

Advancing the craft of technology leadership

DECEMBER 2013

Big Data for Everyone

Hunk™: Splunk Analytics for Hadoop

SPONSORED BY splunk>

CONTENTS

Introduction

For those of us trying to make the most of business data, it used to be a sufficient metaphor to describe the profusion of machine data as "drinking from a fire hose." Even this analogy is no longer adequate. Now, it's more like drinking from a waterfall. One thing is certain: what's in that waterfall is worth a lot more than water—and companies like Splunk have been helping customers pan for proverbial gold for close to a decade now.

More than 6,000 customers have validated Splunk as the leading provider of machine data analytics and operational intelligence. How can Splunk help address today's big data challenges?

If that deluge of data has grown into a waterfall, then Hadoop, the distributed file system, has become the lake beneath it. Companies large and small have turned to Hadoop to store the masses of data they collect—the typical Hadoop cluster contains terabytes or petabytes of data.

Hadoop is great because it stores all kinds of data without an express structure or schema, using commodity hardware. The not-so-great thing about Hadoop is the challenge of analyzing that data once it's in Hadoop or moving it somewhere else for analysis. If some financial firms during the last crisis were deemed "too big to fail," then data in Hadoop is "too big to move."

That's why Splunk came up with Hunk™: Splunk Analytics for Hadoop. Hunk allows mere mortals to interact with and ask questions of huge datasets stored in Hadoop. Hunk's ability to create virtual indexes of raw or partially structured data allows business and IT stakeholders with little training to answer questions and opens up data in Hadoop to a wide audience. But that's just the beginning.

Hunk allows mere mortals to interact with and ask questions of huge datasets stored in Hadoop.

Challenges of Today's Analytical Landscape

Let's examine the challenges that organizations are running into after they set up Hadoop clusters and begin storing data in Hadoop.

Fundamentally, people need to find a way to garner value from the massive amounts of big data that pass through their organizations. But big data presents an inherent obstacle because big data is uneven, disparate, incomplete and often in motion. It's hard to get a grasp of big data in a way that delivers value.

Machine data is a critical subset of big data—it's the fastest growing, most complex and most valuable subset of big data, largely because of its sheer ubiquity. Every GPS device,

RFID tag, interactive voice response (IVR) system, database and sensor—almost anything that uses electricity—generates machine data that can tell companies something important about the way their businesses actually run each day.

Machine data is valuable because it contains records of user behavior: purchasing habits, security violations, fraud attempts, social media posts and customer experiences, for example. Though Hadoop has made machine data easier to store, its value is elusive because few have the time or money to build a "science project" out of Hadoop and develop assorted tools to deliver an effective analytical capability.

Few have the time or money to build a "science project" out of Hadoop and develop assorted tools to deliver an effective analytical capability.

Splunk Enterprise: A Quick Refresher

Splunk has been in the business of extracting value from machine data for nearly a decade. To deal with this situation, Splunk developed Splunk Enterprise, which sifts through machine data to provide analytics in real time for up to hundreds of terabytes a day of streaming and historical data. Splunk Enterprise supports the "four Vs" that characterize big data, and especially machine data:

- **Volume.** Splunk Enterprise accommodates the waterfall of machine data with a scalable, real-time architecture.
- **Velocity.** The Splunk Enterprise architecture addresses the speed and scope of the data flows with an architecture that scales horizontally across commodity hardware. Splunk expands rapidly to meet unanticipated analytical needs.
- **Variety.** One of the essential characteristics of the "bigness" of big data comes from the wide variety of data sources and types. Splunk Enterprise manages forwarding and indexing of highly diverse raw data from thousands of heterogeneous sources.
- **Variability.** Companies collect data voraciously in anticipation of future usefulness. That means they don't need to apply a schema to the data while it's collected. As such, Splunk supports a late-binding schema for analyzing raw, unstructured or polystructured data.

Splunk Enterprise is the industry leading solution for analyzing machine data. But what about analyzing historical data in Hadoop?

Hurdles of Hadoop

Hadoop provides the advantage of storing data cheaply. But when left unmanaged, businesses and the public sector struggle to use it for analytics. Some of the known challenges of Hadoop include:

Cost. Cheap storage has its price for analytics. According to Gartner, those who attempt to create custom applications, or even purchase off-the-shelf applications to wring analytical value from Hadoop, wind up spending as much as 20 times more on services (read: consultants) as they do on software.[1]

> According to Gartner, companies working with Hadoop analytics spend 20 times more on services than on software.

Specialized skills. Getting any kind of analytics out of Hadoop data requires rare, specialized skillsets—at the very least, a mastery of MapReduce, the programming model that processes data stored in the Hadoop Distributed File System (HDFS).

Slow results and no preview of results in progress. MapReduce runs slowly. How much time is lost waiting for a batch job to finish? Queries can take as long as getting a cup of coffee or may run overnight. If the batch job doesn't produce useful results, the process starts all over again. Most businesses don't have that kind of time.

Multi-party landscape. Hadoop is not one thing. It consists of 13 or more open source projects and sub-projects that need integration—and no one entity is in charge of that. Picking a Hadoop distribution such as Cloudera, Hortonworks, IBM, Pivotal or MapR helps, but the knowledge curve needed for keeping track of all the open source projects related to Hadoop is as steep as that required to master MapReduce itself. Most distributions assume users enjoy integration and experimenting. Some do—but do you?

Predefining schemas. To overcome slow MapReduce jobs, the Hadoop community has introduced options for Hive or SQL on Hadoop. These require predefining schemas, which is impossible or impractical given the variability of raw, unstructured, and polystructured data in Hadoop. It also invalidates Hadoop's value proposition, which is that it can easily accept and store data types without pre-definition.

[1] Gartner, Big Data Drives Rapid Changes in Infrastructure and $232 Billion in IT Spending Through 2016, October 12, 2012.

A Schema for the Schema-less: The Problem with SQL on Hadoop

Take a machine data file such as /var/log/messages, which may contain dozens or hundreds of formats. Each format may potentially hold valuable data. If we approach this in the way SQL on Hadoop solutions do, we either:

- Create multiple tables for each data type, which is a significant amount of work or
- Hand-build a very sparse table with all the fields that might be applicable.

Even with JSON or Avro data in Hadoop, each entry may contain a distinct schema.

SQL on Hadoop therefore invalidates the value proposition of storing data without pre-defined schemas.

The question we now must ask is: How do you get value out of data that is "too big to move" without limiting flexibility by attempting to pre-define schemas for data that by its very definition is varied and variable?

What Is Hunk?

In response to these hurdles, Splunk created Hunk™: Splunk Analytics for Hadoop. Hunk is a full-featured, integrated analytics product that aims to deliver actionable insights from raw data. It delivers interactive data exploration, analysis and visualizations for Hadoop, making it much easier to justify a business case for unlocking the value of data stored in Hadoop.

A Sampling of Hunk Use Cases

- Data analytics for new product and service launches
- Synthesis of data from multiple customer touchpoints (IVR, RFID, online purchases, tweets, etc.) for a 360-degree view of the customer
- Comprehensive security analytics to protect against contemporary threats
- Easier application development for big data apps on top of data stored in Hadoop

How Hunk Does It

Hunk's capabilities derive from these key ingredients:

Virtual Index. This capability allows users to leverage the existing Splunk technology stack against data wherever it rests. This includes the Data Model and Pivot Interface Splunk first introduced with Splunk 6.

Schema-on-the-fly. Instead of requiring users to know all the questions they want to ask of data from the start, Hunk allows them to ask and answer questions of data in Hadoop with schema-on-the-fly. The structure of that schema is applied at search time, and it can automatically find patterns and trends. Hunk takes schema-on-the-fly to the furthest extent possible—even things like event breaking are done at search time.

Flexibility and fast time-to-value. Hunk affords flexibility and speed of insights that don't normally come from conventional off-the-shelf products or "science projects." It normalizes data as needed, but not by a predetermined requirement. Its search language has a lot more in common with Google and web browsers than it does with legacy business intelligence platforms. Since it's unlikely two users will have the same question for the same dataset, Hunk also supports multiple views into the same data.

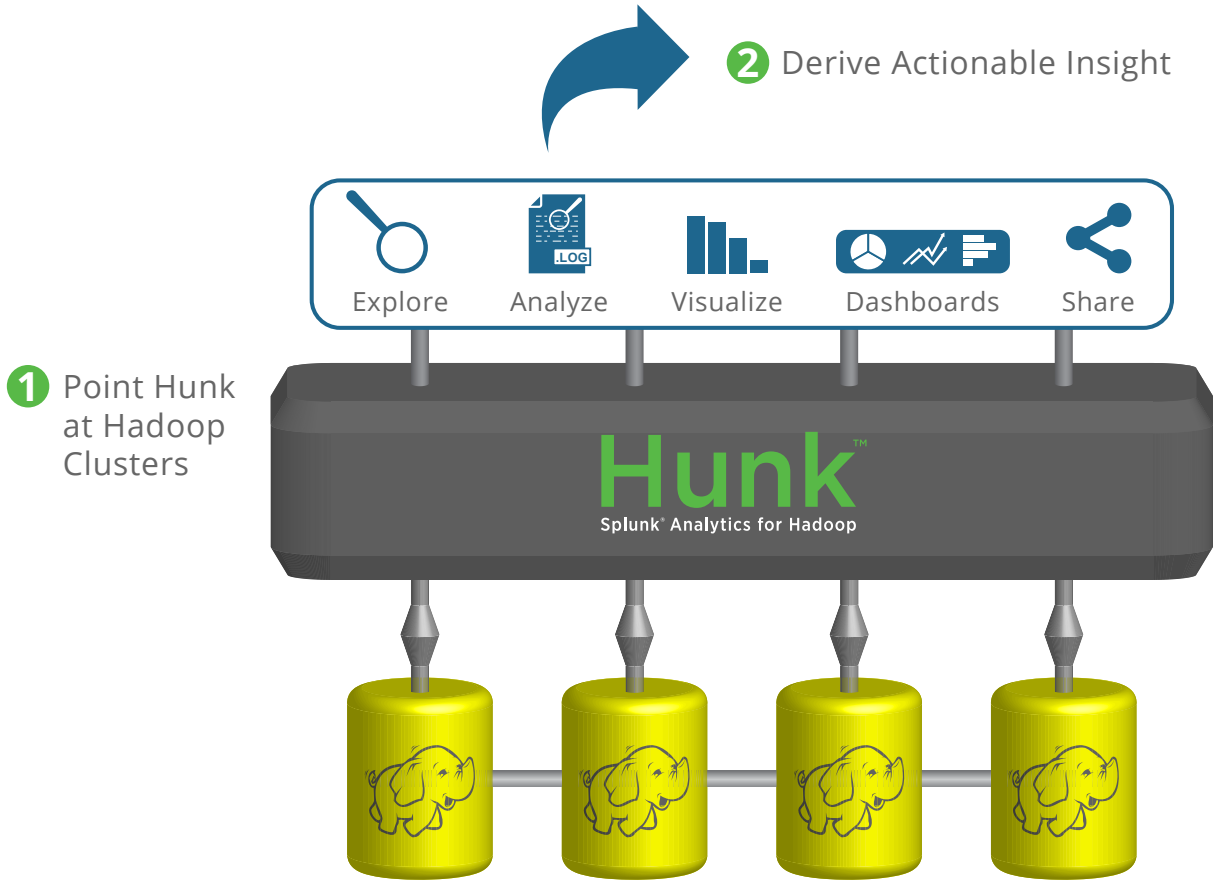
Hunk takes schema-on-the-fly to the furthest extent possible—even things like event breaking are done at search time.

Leaving Hadoop Alone

Some of the analytics tools of the past overcame Hadoop's unwieldy topography by siphoning out small increments of data at a time, breaking it down, analyzing it and then (hopefully) returning it back to the Hadoop file system or an external in-memory store in an improved format. That takes a lot of time. Hunk does not change any of the raw data in HDFS, nor does it move that data into another data store or data mart—that saves time and ensures that you still have all the original raw data in Hadoop, which is important for asking questions you may not have thought of at first.

Point, Shoot, Analyze

Using Hunk is like a point-and-shoot camera for data—just point it at a Hadoop cluster and start exploring, analyzing, and visualizing. Exploration, analysis and reporting all happen with ease based on the proven power of the Splunk Search Processing Language (SPL™) and all of the work done to make that language powerful and easy to use.



Interactive Data Exploration. With Hunk, search is flexible, intuitive and delivers immediate results. There is no requirement to understand the data up front—the point is to understand it by exploring it. Searching and exploration happen in the same interface, and once trends begin to emerge with the data preview feature, they can be iterated across large datasets or even searches across data in multiple Hadoop clusters. Previewing is one of the many unique features of Hunk—alternative approaches require you to wait for MapReduce jobs to finish before you see any results. Or you're forced to pick a small sample dataset, which ruins the value of big datasets for ad hoc exploratory analytics.

Interactive Data Analysis. Hunk supports multiple types of correlation (time, transactions, sub-searches, lookups and joins) and over 100 statistical commands. You can conduct deep analysis and pattern detection for spotting anomalies or new trends in your data.

For example, you can get a 360-degree view of your customer by analyzing operational records, website logs, social media and more. You can address advanced persistent security threats by studying historical data and adding data sources such as packet flows, NetFlow, DNS logs, building entry logs, application logs and employee postings on social media sites. You can go through reams of product and service usage data to optimize offerings and conduct exploratory analysis and A/B tests to evaluate new offerings.

Reporting and Visualization From Hadoop. Instead of sifting through grains of sand, you can generate reports on the fly from difficult-to-understand data. Schedule report delivery for management. Create custom dashboards with multiple charts, views, reports and external data sources, all while enabling you and the stakeholders you support to drill down at any point to the original raw data.

Far from requiring specialized skills and systems, with Hunk, you can personalize and share the data via PDF or view and edit dashboards on any desktop, tablet or mobile device. Hunk offers secure personalization through role-based access controls, an important feature missing from raw Hadoop, which provides access to all the data in Hadoop or none of it.

Alternatives to Hunk

There is always more than one way to accomplish an analytical task. It's a question of audience and emphasis.

The **Do-It-Yourself** approach, using MapReduce or Pig, is for the true Hadoop "ninjas." It's difficult to integrate all of the pieces that make up Hadoop. MapReduce skills are rare and expensive, and jobs on MapReduce can run very slowly—and you don't know what you're getting until they're done. Hunk doesn't require an expert—its visual interface is designed for business analysts and IT users. Ninjas are welcome, but not necessary. Hunk abstracts the complexities of MapReduce, making use of Splunk's search-processing language which is optimized for unstructured or arbitrarily structured data, is naturally interactive and offers a visual interface for analyzing data.

Using Hive or SQL-on-Hadoop appeals to customers because it leverages existing SQL skills. However, this approach forces structure onto naturally unstructured data. Any data that doesn't "fit" gets lost, recreating the problem Hadoop was meant to solve. Further, this approach requires knowledge of the underlying data, even when writing SQL.

Extracting data to an in-memory store has become a popular approach because it doesn't require direct advanced knowledge of Hadoop—just migrate the data out of Hadoop to a separate data mart or in-memory data store. But the problems of Hadoop dog this methodology also. The data is too big to move all at once, and there is limited drilldown. There is no opportunity to preview results and it becomes yet another "data mart" to manage.

Hunk: Hadoop for Everyone

So far, using Hadoop has required experts. Hunk opens up Hadoop to meet the needs of everyone, from line-of-business users to enterprise developers. Business users such as data analysts, product managers and business analysts conduct batch analytics, funnel analysis and long-term reporting. Enterprise developers find Hunk useful because of its API and software developer kits (SDKs) in languages such as Java, JavaScript, Python, PHP, C# and Ruby.

Broadly speaking, Hunk bridges the critical gap between everyday business analysis and Hadoop's idiosyncrasies. It gives broader user groups insight into their data assets without custom development, costly data modeling or lengthy batch process iterations. It works with your data wherever you have it—with the leading distributions, such as Cloudera, Hortonworks, IBM, MapR and Pivotal, as well as downloads from Apache Hadoop.

> *"Hunk gives business analytics teams using Hadoop in their stack an enormous opportunity to improve overall efficiency for everyone."*
>
> Marcus Buda, senior data architect at the Otto Group

Most data management projects are designed to answer a pre-set list of questions, fitting into brittle schemas and a rigid data model. Hunk doesn't have these limitations because the schema is applied at the time of search—so users can immediately ask new questions while they search.

Additionally, Hunk's interactive analytics interface with previews of results dramatically improves the user experience and the speed with which tasks can be accomplished.

> *"I'm super excited about Hunk. Hunk is solving one of the top issues that our customers have: access to the skills and know-how to leverage data in Hadoop. Splunk has a beautiful UI that is very easy to learn. So it bridges that gap and makes it very easy to access data in Hadoop."*
>
> Dr. Amr Awadallah, CTO and co-founder, Cloudera

There's a lot in Hunk for everyone:

- **Business analysts** save time by pointing Hunk at the Hadoop cluster. They can avoid low-level tooling, preview results and answer questions iteratively, without waiting for MapReduce jobs to finish or predefining schemas.
- **Developers** can build scalable enterprise applications based on data in Hadoop, using the developer tools and frameworks they already know.
- **IT managers** can empower users to access and benefit from Hadoop data without going through data "gatekeepers," which creates a queue for scarce resources to write MapReduce jobs. IT departments can provide users with a platform to explore, analyze and visualize data in Hadoop.
- **Data scientists** can democratize and evangelize data by enabling a broader group of line-of-business and departmental colleagues to use and benefit from analytics.
- **Data architects** will find that Hadoop fits seamlessly into their enterprise data architecture, as it is much easier to adapt their architecture for big data and to enforce granular security controls by role and group.

Key Features of Hunk

- All levels of users
- Free form data exploration
- Preview search results
- Schema-on-the-fly
- Splunk search interface
- Role-based access to Hadoop data
- Visualization, dashboards and reporting

Hunk Approach Means

- No moving data out of Hadoop
- No MapReduce programming required
- No low-level tooling
- No waiting for MapReduce jobs to finish
- No predefining schemas

Conclusion

CITO Research finds that Hunk fills a critical gap between the in-the-weeds "expert" approach to operating on data in Hadoop or extracting it for quarantine in an additional system that requires its own skill set and resources.

With Hunk, businesses can rapidly explore, analyze, visualize and share data in Hadoop, without worrying about the vagaries of Hadoop itself. They can easily create custom dashboards for different users and roles. Businesses can protect data with secure, role-based access controls. Through Hunk, the value of Splunk software is opened to an entirely new audience of Hadoop users—which, given the unending and increasing volume of data flowing over the falls, is a group that is getting larger every day.

This paper was created by CITO Research and sponsored by Splunk.