BIG DATA IN THE CLOUD
SUCCESS SHEET

Qubole

A recent study **from Capgemini** found that only a fourth (27 percent) of executives say their respective in-house big data initiatives were successful. Such a low rate can be attributed to failing to identify the best infrastructure to be able to scale the project on-demand. However, even adopting a scalable "big data in the cloud" service requires careful planning and consideration. This success sheet details critical factors to consider when selecting and working with a "big data in the cloud" vendor.

ONE:

FULL ORGANIZATIONAL SCALABILITY

Average Qubole customer cluster scales up to 34x times

It is well-known that scalability is a key benefit of storing and processing data in the cloud. Unfortunately, scaling Hadoop workloads can be very difficult. These workloads are bursty by nature, with CPUs often maxing out for a few minutes before becoming idle. Constant monitoring of node utilization is not a simple or scalable way to maximize efficiency. Due to this, big data solutions will require varying levels of management to scale a cluster up or down to meet load demand and won't offer true auto-scaling capabilities. To learn how Qubole was able to overcome these roadblocks and offer auto-scaling Hadoop technology, see this **blog post.**

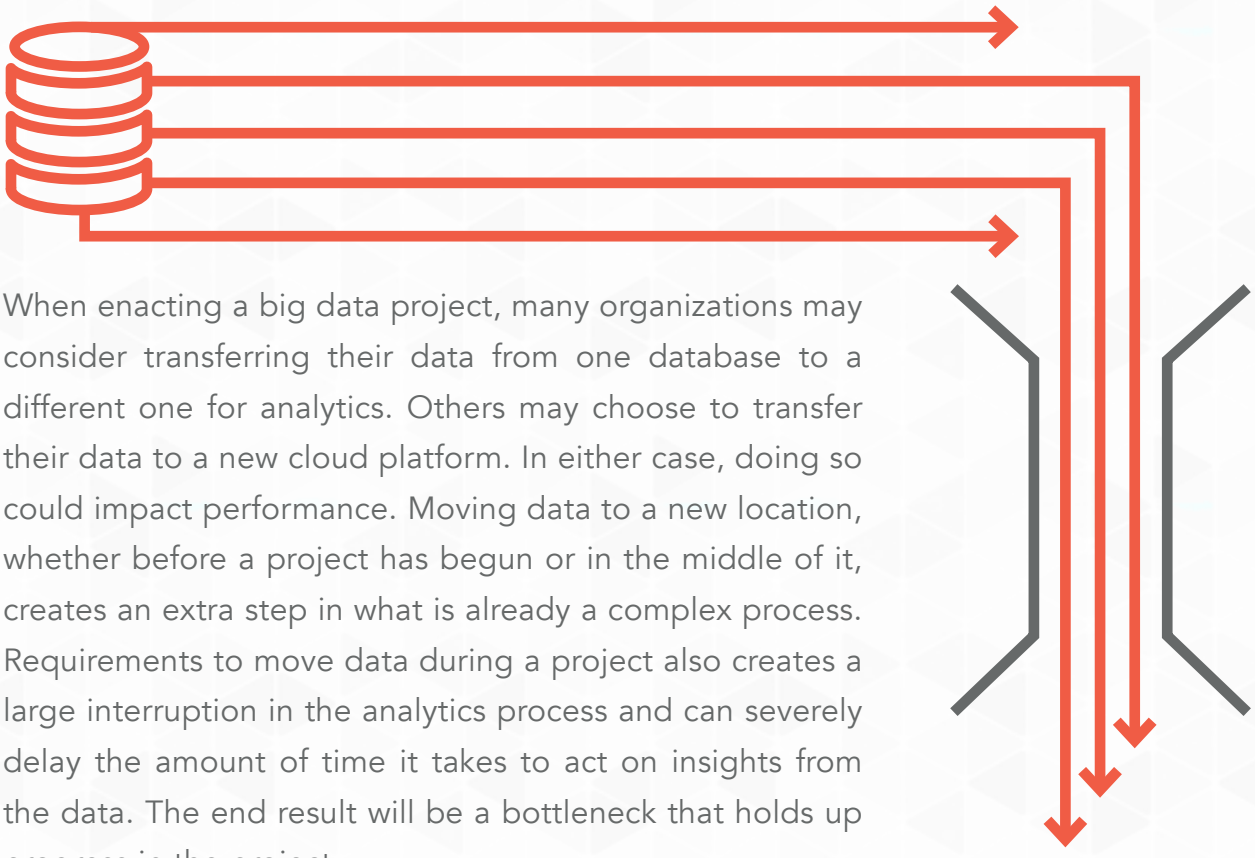
In addition to technological scalability, stakeholders should also consider organizational scalability. A big data initiative's true success lies not in the technology but in the actionable insights businesses are able to take away from the data. The more users who are able to access the platform without administrative support, the more efficiently the organization will be able to meet its true business objectives. To meet this requirement, businesses should identify a vendor with a self-service model, driven by policy-based automation. Additionally, leaders should invest in developing structure and culture for a data science team to work together effectively.

For every 1 administrator, there are 21 users that are directly accessing data through Qubole's self-service platform.

TWO

DON'T MOVE YOUR DATA

Qubole Data Service accesses, computes and persists data in the customer's' owned and controlled cloud accounts.



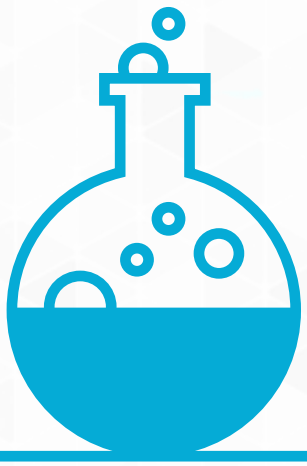
When enacting a big data project, many organizations may consider transferring their data from one database to a different one for analytics. Others may choose to transfer their data to a new cloud platform. In either case, doing so could impact performance. Moving data to a new location, whether before a project has begun or in the middle of it, creates an extra step in what is already a complex process. Requirements to move data during a project also creates a large interruption in the analytics process and can severely delay the amount of time it takes to act on insights from the data. The end result will be a bottleneck that holds up progress in the project.

THREE

DATA SCIENCE FUNDAMENTALS

Many companies focus on collecting as much data as possible from as many sources as possible. While gathering data is important, the second half of the equation — the "science" part — is too often forgotten. You need to approach your big data efforts from a scientific perspective to gain the most benefit from them. If not, you're at risk of basing your decisions off of bad models, poor data quality, and erroneous assumptions. Focus on making sure your measurements are accurate and your biases are eliminated. Incorporating the science with the data is a much wiser move to make.

Data science can become a complicated endeavor regardless of your type of business. If data teams are unleashed to find whatever insights they want, they'll quickly end up wasting resources in a directionless effort. Companies that are successful at data science have established an oversight body for all data governance, ensuring each data team is getting the right data and that they're working with it in a way that satisfies business objectives. They also make sure the entire organization's data processes are operative and progressing in just the right way.



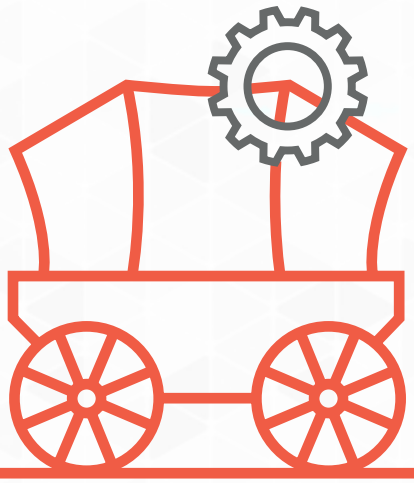
FOUR

HAVE AN OBJECTIVE

Many businesses hear that big data analytics is valuable for business decision making, so they jump on the bandwagon without much thought. Collecting tons of data without an end goal established will only make for confusion and frustration down the road. Create goals that allow you to measure your progress along the way. You'll also need to take into account what data you need, what existing data you have, and how it all applies to your business objectives.

The majority of big data projects fall into one of two broad categories: storage driven or app driven.

For organizations running corporate internal applications in a private data center now faced with large volumes of structured and unstructured data, an in-house Hadoop infrastructure is well suited to what would primarily be data-storage driven projects, especially for organizations where broad accessibility and frequent use are not necessary.



Collecting information into a data lake is one thing, but finding the business value hidden in heaps of structured and unstructured data is quite another. The next generation of big data applications are designed with the business user in mind. They should be simple to use, easy to navigate and they give organizations complete internal control of their data. Unlike on-premises or older, developer-focused Hadoop solutions where big data analysis for insights is left up to data scientists with teams of supporting engineers, the next generation of big data applications put analysis and insight discovery into the hands of business users—the people who understand the company's core business strategies and objectives. Thanks to the ease of data accessibility that apps provide through a user-friendly interface, execs and analysts on both the tech and business sides can work together to solve real business problems across the enterprise, making sure that the big data strategy always serves the strategies and objectives of the business.

FIVE:

CORE REQUIREMENTS SUCH AS SECURITY AND RELIABILITY

Security has long been a concern for those considering cloud adoption. Yet, the numbers show that on-premises data centers face just as many threats as cloud environments. **From 2012 to 2013**, vulnerability scanning attacks increased from 28 percent to 40 percent for on-premises data centers and increased from 27 percent to 44 percent for cloud-hosted environments. Infrastructure as a Service (IaaS) also offers the benefit of automatic updates or patching and the ability to take advantage of the latest security tools, such as VPC and identity-based access controls.

Uptime and data availability are also important to consider. Rather than putting ops resources toward maintaining upkeep and availability of systems and data, organizations can rely on the cloud provider's reliability. Both Amazon Web Services and Google Cloud Platform boast four nines of availability and up to 11 nines of durability for storage.

Integrate with cloud providers' security measures - Start by integrating your big data platform with the encryption and security practices inherent in your cloud environment. Select a vendor that offers a centralized control panel to manage data governance.

Don't overlook basic security measures – To ensure user identification and control user access to sensitive data, it's important to create users and groups and then map users to groups. Assign and lock down permissions by groups, and impose the use of strong passwords. Organizations can take further steps to minimize bad passwords by opting for Single Sign-On (SSO) when setting up new accounts. SAML and other SSO technologies can help simplify accessing tools for employees and makes it easy for administrators to add and remove employee access to company data. Assign fine grained permissions on a need-to-know basis only, and avoid broad stroke permissions as much as possible.

Choose the right remediation technique – When business analytic needs require access to real data, as opposed to desensitized data, there are two remediation techniques to choose from—encryption or masking. While masking offers the most secure remediation, encryption might be a better choice as it offers greater flexibility to meet evolving needs. Either way it's important to ensure that the data protection solutions being considered are capable of supporting both remediation techniques. That way, both masked and unmasked versions of sensitive data can be kept in separate file directories if desired.

Ensure that encryption integrates with access control – Once chosen, an encryption solution must be made compatible with the organization's access control technology. Otherwise, users with different credentials won't have the appropriate, selective access to sensitive data in the Big Data environment that they need.

Monitor, detect and resolve issues – Even the best security models will be found wanting without the capability to detect non-compliance issues and suspected or actual security breaches and quickly resolve them. Organizations need to make sure that best practice monitoring, and detection processes are in place.

Ensure proper training and enforcement – To be fully effective, best practice policies and procedures on data security in Hadoop and other systems must be frequently revisited in employee trainings and constantly supervised and enforced.

SIX:

OPTIMIZE USE OF THE OBJECT STORE

There are challenges to making some aspects of the cloud work well with existing Hadoop and Spark software. In particular, object stores, despite the benefits of possessing nearly unlimited capacity and scalability, don't always work like file systems, which is an expectation for Hadoop Distributed File System (HDFS). In particular, it can be challenging to optimize performance on an object store such as Amazon S3, where directory listings and move operations are not very performant. A successful Hadoop deployment will need to bridge the performance gap between the object store and HDFS. To learn more about Qubole's optimization techniques around object storage, see this **blog post.**

SUMMARY

- **Full Organizational Scalability:** Ensure your project is fully scalable at the technology and team level.
- **Don't Move Your Data:** Avoid roadblocks by keeping your data in its original data store.
- **Data Science Fundamentals:** Remember the basics of clean data, proven data science models and overall data governance.
- **Set a Business Objective:** Big data is not an IT project. Have short and long-term goals in mind.
- **Security Best Practices:** Remember remediation techniques, encryption and proper enforcement.
- **Optimize Use of Object Store:** Separating storage and compute can impact performance, so remember to optimize for these unique challenges.

ARE YOU READY TO GIVE QUBOLE A TEST DRIVE?

SIGN UP FOR A RISK-FREE TRIAL

BEGIN FREE TRIAL

Qubole