Running Big Data Infrastructure: Five Areas That Need Your Attention

When running a Big Data infrastructure, focus on five key areas will ensure the right choices are made for a successful deployment.

Qubole

Table of Contents

Qubole

Big Data is One of the Most Talked about Topics in IT

Qubole

That’s no surprise.

Businesses quickly discovered that leveraging data allows them to better compete in the marketplace. But, in the rush towards Big Data adoption, most conversation focuses on applications, not the infrastructure that supports any successful deployment of Big Data. While the business and IT communities debate the merits and uses of Big Data, data engineers are quietly focused on architecting and running infrastructures that are efficient, cost-effective, and scalable.

Qubole

The Importance of Infrastructure to Big Data Can't Be Overstated.

A building's architecture ultimately determines whether it stands or falls. In the same way, the infrastructure supporting Big Data applications can make or break the company relying on the data to inform critical business decisions.

The most visible and dramatic impact of poor infrastructure is a complete collapse, resulting in loss of service to the users depending on the data. The jarring effect is similar to suddenly shutting off the lights – everyone is going to notice. This type of failure is both extremely costly and time-consuming to remedy.

Another danger is a slow financial drain on the company. Poorly designed infrastructure is expensive to maintain. Once the wrong technology choices have been made, the costs to keep the infrastructure functional begin to snowball as the volume, variety, and demand for data increase.

Finally, the worst outcome of a flawed infrastructure is drawing false conclusions. Because the company is unaware of the imperfections in the infrastructure and resulting flawed data, critical business decisions are made based on bad information. These decisions have a direct impact on product development, customer service, security, and ultimately, revenue.

With the stakes this high, data engineers are under tremendous pressure to make the right choices when architecting Big Data infrastructure.

New Challenges Arise Following Big Data Adoption.

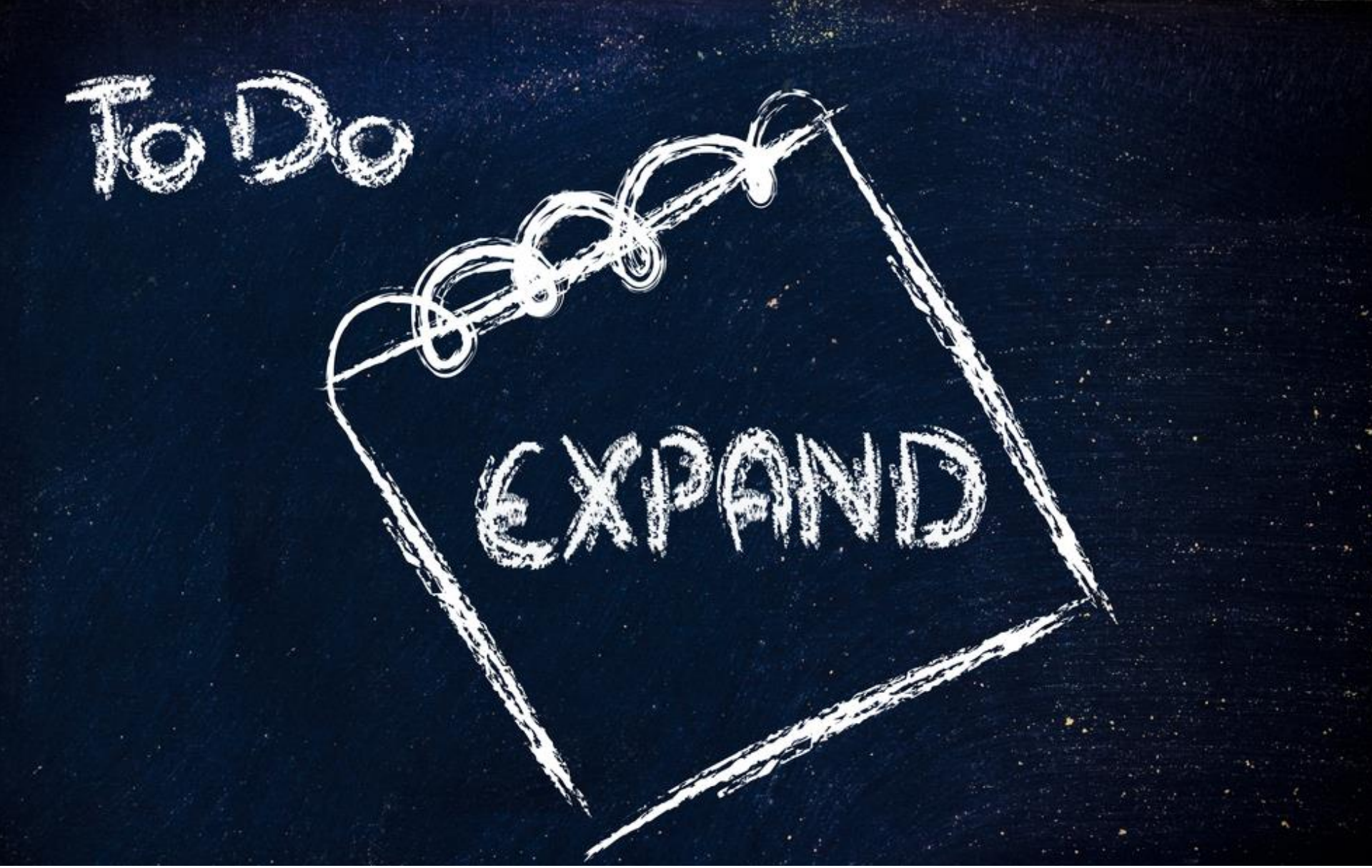
For companies that have already bought into Big Data and developed a use case, the focus shifts to using data more efficiently. Data engineers responsible for Big Data infrastructure face three big challenges as a result of this change in perspective.

Controlling Costs
The unpredictable nature of data can result in not only headaches but also runaway costs. Data engineers seek ways to run infrastructure in a controlled manner to improve and contain costs.

Scale
There is no doubt that data usage will increase. The uncertainty lies in when, by how much, and for what duration. Data engineers need an infrastructure that can react to variation in usage patterns.

Management
While supporting the infrastructure could be a 24/7 job, data engineers seldom have the time or staff to dedicate to this task. A well-designed infrastructure allows the data engineer to step away to manage day-to-day operations and increase productivity elsewhere in the department.

Qubole



Qubole

Five Areas That Need Your Attention

Five Areas That Need Your Attention

Designing a Big Data infrastructure that meets these challenges requires careful consideration in 5 key areas.

1. Elasticity

Scalability generally refers to the ability of a technology to support increasing demand. Typically, in order to scale infrastructure, more machines are added. Elasticity refers to the ability of a technology to flex up or down depending on demand at any given time. Working on an elastic substrate allows a company to utilize one machine for lots of operations, not just Big Data, without affecting the budget.

2. Reliability

Business users count on the availability of data at regular intervals to efficiently run the company. Reliability begins with how effectively data enters the infrastructure and ends with predictable data delivery. Companies, such as Facebook, utilize internal service level agreements (SLAs) to ensure data availability across teams.

Qubole

Five Areas That Need Your Attention

3. Self-Service Tools

Business analysts require data to perform their job functions, yet without a deep technical skill set, quickly and easily accessing the required data can be practically impossible. In the past, companies resolved this issue by hiring teams to act as liaisons between the data engineers and business users. Today, self-service tools provide ready access to data through a user-friendly interface. Productivity is increased across the board making it possible for more people in the organization to make data-driven decisions and making it possible for data teams to support larger and larger numbers of business users.

4. Monitoring

Active monitoring of the infrastructure minimizes issues and maximizes availability. How much monitoring? The more, the better. With the introduction of self-service tools, more users relying on the data will be impacted by inefficiencies or other concerns. Monitoring provides insight into how the system is operating and quickly identifies minor glitches that can be corrected before they develop into major problems.

5. Open Source

Open source technology evolves rapidly, and new features, such as faster queries, could result in increased productivity, improve cluster availability, or even competitive advantage in the marketplace. Data engineers must stay on top of the latest versions to ensure that the infrastructure runs at peak performance. Working with a knowledgeable vendor can help data engineers mitigate the risk of missing even one update.

Five Areas That Need Your Attention

In the early days of Information Technology, an investment in computers and software could be a significant business advantage by improving overall operations. The same thing is happening today as Big Data provides detailed insight into customer behavior, product usage, and more.

A well-crafted data infrastructure means that analysts can access this Big Data framework to come up with good models to grow the company, reduce costs, and innovate. Making the right choices in these 5 key areas will allow data engineers to develop an infrastructure that controls costs, scales, and remains low maintenance while delivering on the promise of monetized data.

Qubole

About the Authors

Ashish Thusoo

Co-Founder and CEO, Qubole Inc.

Ashish has a robust history in data engineering, beginning his career at Oracle as a contributor to

the RDBMS product.

As head of the data infrastructure team at

Facebook, Ashish created one of the largest data processing and analytics platforms in the world.

Many of the artifacts developed for this revolutionary deployment have become mainstream to Big Data implementation.

Ashish co-create Apache Hive and served as the project's founding vice president at the Apache Software Foundation.

He holds a bachelor's degree in computer science from IIT-Delhi and a master's degree from University of Wisconsin-Madison.

Qubole

About the Authors

Joydeep Sen Sarma
Co-Founder & CTO, Qubole Inc.

Prior to co-founding Qubole, Joydeep worked at Facebook where he boot-strapped the data processing ecosystem based on Hadoop, started the Apache Hive project, and led the Data Infrastructure team. Joydeep was a key contributor on the Facebook Messages architecture team that brought Apache HBase to Facebook and to the transactional and reporting backends for Facebook Credits.

He cut his teeth building data driven applications as the lead engineer on Yahoo's in-house Recommendation Platform. Joydeep holds numerous patents, has many published papers, and has been both speaker and panelist at Hadoop summits and at other Silicon Valley conferences.

Joydeep studied computer science at IIT-Delhi and University of Pittsburgh and started his career working on Oracle's database kernel and building highly available and scalable file systems at Netapp.

Qubole

Get help running your big data infrastructure

Learn how Qubole delivers the next generation Big Data Service for data professionals who perform data integration, exploration, analysis and job scheduling in the cloud.

Sign up for Qubole's 15-day trial.

Info.qubole.com/free-account

Qubole

About Qubole

Qubole delivers the next generation big data service for data professionals who perform data integration, exploration, analysis, and job scheduling in the cloud.

Features:

100% managed hadoop cluster in the cloud (auto-scaling, high performance, self-maintenance) built-in data connectors to apps and data sources (appnexus, redshift, mongodb and more)

24/7 customer support (via chat, phone or e-mail) from our data infrastructure SWAT team deliver more value to your business by focusing on your data and queries, not worrying about provisioning clusters, machines, or job flows.

Start your risk-free 15-day trial instantly. No credit card required.