How to Do a Big Data Project

Big Data is sweeping the business world – and while it can mean different things to different people, one thing always rings true: Data-driven decisions and applications create immense value by utilizing data sources to discover, present, and operationalize important business insights.

While there is broad industry consensus on the value of Big Data, there is no standardized approach for how to begin and complete a project. The many tools, vendors, and trends in the marketplace, multiplied by different use cases and potential projects, can lead to decision paralysis. In addition, some companies mistakenly focus on technology first instead of business objectives.

At Infochimps, we want to share our experiences after guiding many enterprises through successful Big Data projects. This project guideline should empower you to tackle the discussion and decide on build versus buy when it comes to achieving your defined business objectives across various technical environments.

All of these factors put every Big Data project at risk. In a *recent survey of IT professionals*, it was found that nearly 55% of Big Data projects don't get completed. The same metric for IT projects in general, is only 25%.

While how you manage your Big Data project will vary depending on your specific use case and company profile, there are 4 key steps to successfully implement a Big Data project:

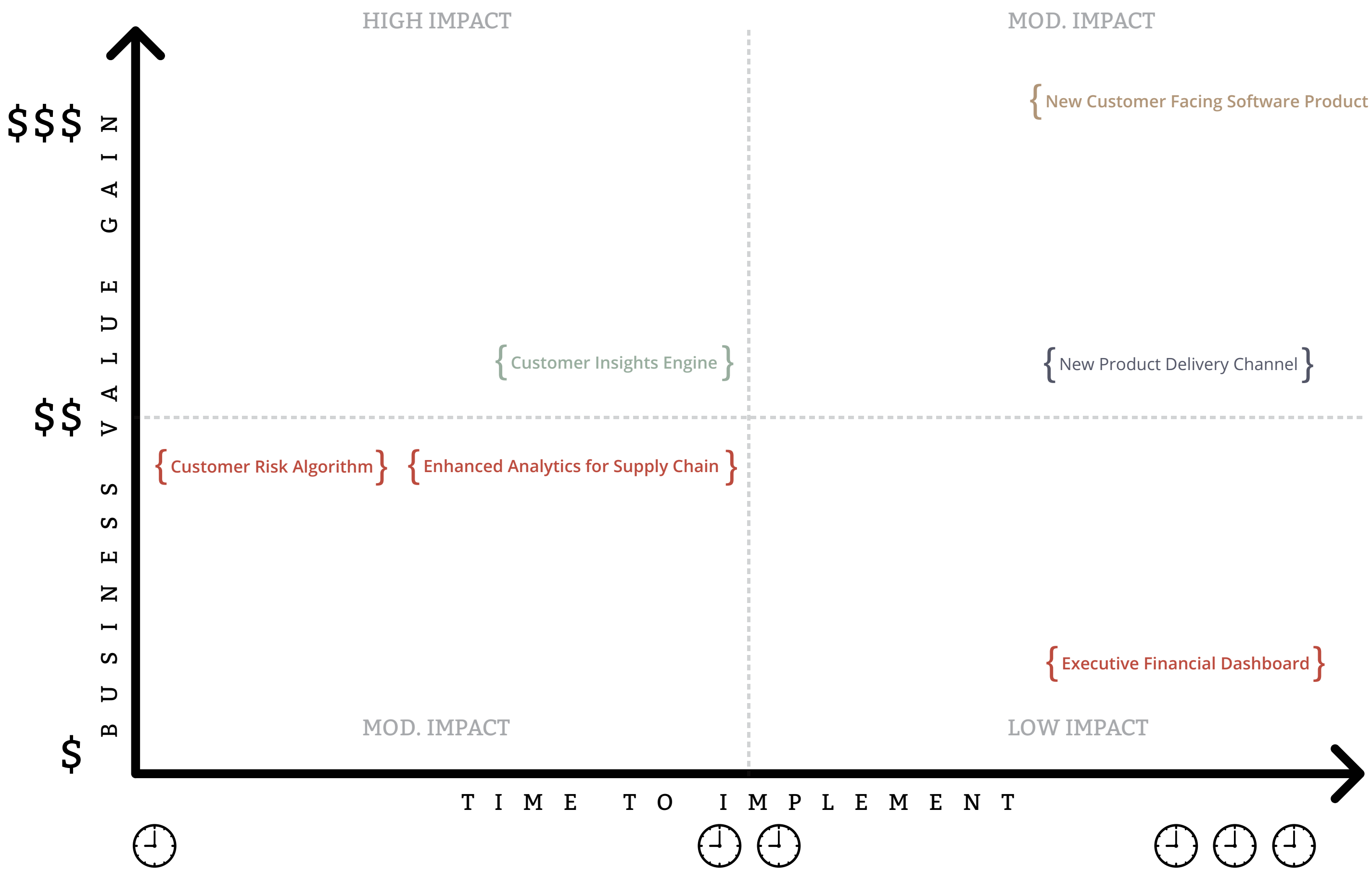
1. **Defining your business use case** - with clearly defined objectives driving business value.
2. **Planning your project** - a well managed plan and scope will lead to success.
3. **Defining your technical requirements** - detailed requirements will ensure you build what you need to reach your objectives.
4. **Creating a "Total Business Value Assessment"** - a holistic solution comparison will take the politics (and emotion) out of the choices.

1. Define Your Business Use Case

As enterprises explore Big Data, the business drivers vary widely from revenue growth to market differentiation. We've seen companies realize the most significant benefits from Big Data projects when they start with an inventory of business challenges and goals and quickly narrow them down to those expected to provide the highest return.

> **At Infochimps**, we have seen Big Data business objectives ranging from "creating and deploying a SaaS application product" to "increasing revenue by providing the sales team prioritized leads leveraging customer service data."

Typical Use Cases



🟫 Revenue Growth 🟩 Product/Service Quality ⬛ Market Differentiation 🟥 Cost & Risk Mitigation

{Enhanced Analytics for Supply Chain}	$$	🕒🕒	{Customer Risk Algorithm}	$$	🕒
{Executive Financial Dashboard}	$	🕒🕒🕒	{Customer Insights Engine}	$$	🕒🕒🕒
{New Product Delivery Channel}	$$	🕒🕒🕒	{New Customer Facing Software Product}	$$$	🕒🕒🕒

In order to explore your organization's expectations of Big Data, we recommend answering these questions first:

- What is the project goal?
- What direction is the business headed?
- What are the obstacles to getting there?
- Who is are the key stakeholders and what are their roles?
- What is the first Big Data use case determined by key stakeholders?

These questions build a good foundation for your project. The more specific and connected to your business goals your answers are, the more likely your project will succeed. These are the types of discussion points we ask our clients when first learning about their Big Data projects. Here are more specific questions that elaborate on each foundational point.

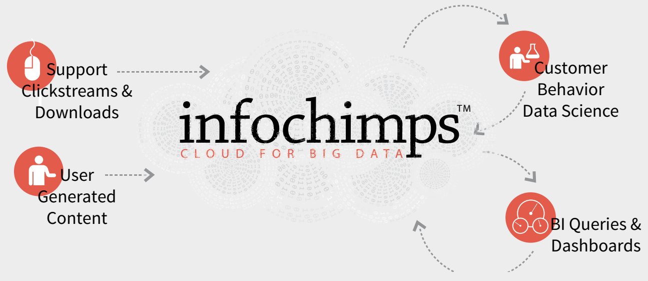
What Direction is the Business Headed?

- Determine the company's high level objectives and how Big Data can support these objectives.
- Understand the company's perception of Big Data and any relative historical context.
- Identify the problem area, such as marketing, customer care, or business development, and the motivations behind the project.
- Describe the problem and obstacles in non-technical terms.
- Inventory any solutions and tools currently used to address the business problem.
- Weigh the advantages and disadvantages of the current solutions.
- Navigate the process for initiating new projects and implementing solutions.

Identify Stakeholders and Business Use Case

- Identify the stakeholders that will benefit from the Big Data project.
- Interview individual stakeholders to determine their project goals and concerns.
- Document specific business objectives decided upon by key decision makers.
- Assign priorities to the business objectives
- Architect a business use case.

At Infochimps, we create these types of business use cases:



Our company intends to generate insights on how customers interact with their customer service portal and improve their services for happier customers and more streamlined business.

→ Faster path to ROI with both tech and services

→ Ability to prove the value of Big Data internally

→ Scalability to more data sources and use cases

Determine the Project Team

- Identify the project “sponsor” to remove obstacles, find the budget, provide organizational support, and champion the cause.
- Establish the project manager and the team. Define the roles and responsibilities of each team member.
- Understand the team’s availability and resource constraints for the project.

> **At Infochimps**, we typically see project teams that include these roles: Executive Sponsor, Technical Sponsor, Project Manager, Architect, Data Engineer, Data Scientist/Analyst, Lead Test, Lead QA, Lead Developer, IT Lead.

Example Business Use Case Checklist:

Item	XYZ Bank
Executive Sponsor:	SVP IT
Company Objective:	Enhance internal audit & reporting capabilities
Pilot (Yes/No - how long):	Yes (3 months)
Budget (Pilot/Production):	$100,000 (Pilot) - $500,000 (Production)
Project Lead:	Senior Architect
Use Case:	Information security, account security, ID theft monitoring and prevention
Success Criteria:	faster reporting, more robust reporting, improved operation efficiency

2. Plan Your Project

This is where things get specific. As a result of your research and meetings, you most likely have a nebulous objective, like “reducing customer churn.” This section intends to construct a concrete and specific objective agreed upon by the project sponsors and stakeholders.

→ **Specify expected goals in measurable business terms.**

→ **Identify all business questions as precisely as possible.**

→ **Determine any other quantifiable business requirements.**

→ **Define what a successful Big Data implementation would look like.**

The goal may now be clear, but how will you know once you’ve achieved it? It’s important to define what business success for your Big Data project looks like before proceeding further.

At Infochimps, we like to see an objective like "Leveraging data from our CRM, Customer Support and Finance applications and using cloud architecture, create an application to score customers on a ranking scale, based on the likelihood we'll lose their business. The app will be used by the Account Services team, who will employ a special customer service strategy to increase retention, thereby reducing churn."

Set Specific Objective Success Criteria

When determining success criteria, it's important to pick criteria that are measurable, such as a specific key performance metric.

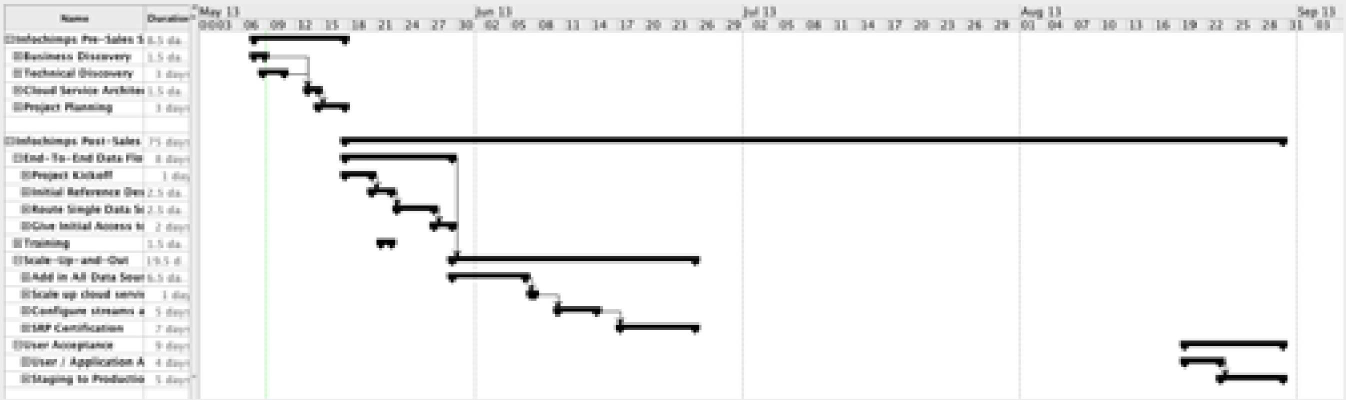
The following are tasks and considerations you can use to ensure you have properly captured the success criteria:

- As precisely as possible, document the success criteria for this project.
- Make sure each identified business objective has a measurable criteria that will determine if that objective has been met successfully.
- Share and gain approval of your business success criteria among key stakeholders.

Tie the Project Plan to Success

- Determine proper scope, specifically what is included and what is not included.
- Develop a rough budget.
- Set a timeline and successful milestones at 3 months, 6 months, and a year.

At Infochimps, we use a Gantt chart template as a starting point for Big Data implementations:



Name	Duration
Infochimps Pre-Sales S	6.5 da.
Business Discovery	1.5 da.
Technical Discovery	3 days
Cloud Service Architec	1.5 da.
Project Planning	3 days
Infochimps Post-Sales	75 days
End-To-End Data Flo	8 days
Project Kickoff	1 day
Initial Reference Des	2.5 da.
Route Single Data So	2.5 da.
Give Initial Access to	2 days
Training	1.5 da.
Scale-Up-and-Out	19.5 d.
Add in All Data Sour	6.5 da.
Scale up cloud servic	1 day
Configure streams a	5 days
SRP Certification	7 days
User Acceptance	9 days
User / Application A	4 days
Staging to Productio	5 days

It is essential to avoid common pitfalls like scope creep, unclear goals, passive or non-existent project management, poor communication, starting too small, etc.

> ***Why do Big Data projects fail 30% more often than other IT projects?***
>
> By far, the biggest reason why these projects failed was inaccurate scope. Requirements expanded out of proportion or there wasn't a firm set of project objectives. Without having firm success criterion, projects continue to expand without realigned timelines, failed to demonstrate positive return on investment, or in the worst case -- failed to meet objectives and provide business value.
>
> Another key point that came up was lack of cooperation between departments. By nature, Big Data is not just going to help one stakeholder. It's often something that improves performance for a lot of different parties such as your IT team, your application developers, your data scientists and analysts, and people across the organization from line of business managers to executives.

For more on this report, see: *http://www.infochimps.com/resources/white-papers/cios-big-data*

3. Define Your Technical Requirements

The technical requirements phase involves taking a closer look at the data available for your Big Data project. This step will enable you to determine the quality of your data and describe the results of these steps in the project documentation.

Current Technical Environment

It's important to understand what tools are used and the architecture they are used in, as it sits today.

- Inventory all tools used today.
- Sketch the current architecture.

Identify Data Sources

Consider what data sources you'll need to take advantage of. Most data that's relevant to a production application's use is going to come from a live stream, or feed:

→ ***Existing data sources.*** This includes a wide variety of data, such as transactional data, survey data, Web logs, etc. Consider whether your existing data sources are enough to meet your needs.

→ ***Purchased data sources.*** Does your organization use supplemental data, such as demographics? If not, consider whether something like a Gnip or Datasift social media and news stream would complement your current data to create additional project value.

→ ***Additional data sources.*** If the above sources don't meet your needs, you may need to conduct surveys or begin additional tracking to supplement your existing data stores.

When examining your data sources ask:

- Which attributes from the database(s) seem most promising?
- Which attributes seem irrelevant and can be excluded?
- Is there enough data to draw generalizable conclusions or make accurate predictions?
- Are there too many attributes for your analytics method of choice?
- Are you merging various data sources? If so, are there areas that might pose a problem when merging?
- Have you considered how missing values are handled in each of your data sources?

There are many ways to describe data, but most descriptions focus on the quantity and quality of the data. Listed below are some key characteristics to address when describing data:

→ ***Volume of data.*** For most analytical techniques, there are trade-offs associated with data size. Large data sets can produce more accurate models, but they can also increase processing time.

→ ***Velocity of data.*** There are also trade-offs associated with whether the data is at rest or in-motion (static or real-time). Velocity translates into how fast the data is created within any given period of time.

→ ***Variety of data.*** Data can take a variety of formats, such as numeric, categorical (string), or Boolean (true/false). Paying attention to value type can prevent problems during later analytics. Frequently, values in the database are representations of characteristics such as gender or product type. For example, one data set may use M and F to represent male and female, while another may use the numeric values 1 and 2. Note any conflicting schemes in the data.

→ ***Time to action.*** Data can be used to take immediate action, as well as be stored for future, non-time critical analysis. It's important to identify which data will most likely be used for real-time actions (<150ms), near real-time actions (seconds), or non-time critical actions (minutes to hours).

At Infochimps, we utilize the following template to gather and compile data sources, documenting each of the dimensions of a data source.

Name	Example
Data Source *Description of the data source*	POS System
Listener Type *HTTP Req./Stream, Syslog, Batch Upload, etc.*	HTTP Request
Data Fields (Attributes)	timestamp, item, purchase price, item ID, inventory ID, purchase ID, customer ID, cashier ID
Data Type *JSON, XML, CSV/TSV/Delimited, Fixed Width, SQL, Binary*	JSON
# of Channels *Either 1 or many; Gnip is 1 source with possibly 100s of diff. URLs to listen to*	1
Data Velocity *Avg. Events / Sec**	125
Max Sustained Events / Sec (10 min)	500
12-Month Growth Factor* *How much is volume expected to grow in next 12-mos.*	5x
Avg Event Size	0.4KB
Time to Action *Real-time actions (<150ms), near real-time actions (seconds), or non-time critical actions (minutes to hours)*	Non Time Critical, by Minute

Identify How You'll Work with the Data

Consider what interfaces and tools are necessary for your company to work with your data sources. Infochimps provides the ability to create custom applications and analytics using native APIs and tools as part of Hadoop, databases, and stream processing – as well as abstracted and unified interfaces for improved user experience. Infochimps also provides users with the ability to produce tables, charts, and other visualization elements using BI tools, such as: Business Objects, Microstrategy, Cognos, Tableau, Datameer, or other similar tools. Such visual analyses can help to address the Big Data project goals defined during the business understanding phase. Other times, it is more appropriate to utilize statistical tools (R, SAS, SPSS, Matlab, etc.) and packaged applications (CRM, POS, ERP, etc.).

- Who needs to work with the data?
- What are their skills and techniques?
- Will training be required?
- What tools do you currently have in your enterprise that you'd like to take advantage of?
- Do those tools have Big Data connectors or proven interface methods?
- What new tools could help with your data mining, analysis, visualization, reporting, etc.?
- How and where will the data be stored?
- What are the reporting and visualization tools necessary to achieve success in your end users' eyes?

> **At Infochimps**, we recommend tracking how you plan to use and consume the data generated by your data sources. Your data sources will eventually "feed" other processes in your Big Data environment (e.g. directly to a customer application, to a RDBMS powering a BI tool, a NoSQL data store, Hadoop, data archive, etc.)

Name	Example
End User *Data Scientist, Data Engineer Data Analyst, Business Analyst, Statistician, BI Specialist, LOB User, Field Manager, Executive, etc*	Data Analyst
End User Tool *Hive/Hue, SQL Server / MySQL / Oracle, Tableau, Cognos, Microstrategy, SAP Business Objects, SAS, SPSS, R, Excel, custom application, custom dashboard, etc.*	Tableau
Analysis Activities *What is the end user doing as part of their analysis?*	Exploratory queries, simple data mashups and calculations, creating visual reports with Tableau's report builder

4. Create a Total Business Value Assessment

Evaluate your options with a "Total Business Value Assessment". This means that you perform at least a 3-year total cost of ownership analysis, but you also include things like time-to-business value, ease-of-use, scalability, standards-based, and enterprise readiness. However, before you get started on evaluating your solution options, it is important to know your "buying team". Buying teams generally consist of stakeholders from multiple organizational levels and sometimes multiple divisions outside of IT. At a minimum, there should be an executive sponsor, project champion or project team lead, technical decision maker, and an economic decision maker.

Your entire buying team needs to be involved in evaluating the options. Options start with who you're relying on to implement your project such as: doing it yourself with internal resources, and/or leveraging software vendors, working with system integrators, deploying with cloud services providers, or using emerging boutique Big Data consulting firms. We recommend looking at each implementation option and weighing them against your specific business priorities. But don't forget to include: Time to Business Value, Ease of Use, Scalability, Standards-based, and Enterprise Readiness. As you evaluate solutions, document how each solution performs on these and other important dimensions.

Time to Business Value. As projects require significant upfront investment, it is important to understand how long before the solution will start generating value. As many of these projects can overextend vendor agreed timelines, it is recommended to request information on similar scope projects that have already been completed for other clients to better determine whether the vendor delivers on set milestones. If your new Big Data application generates $5M per month and you launch 12 months ahead of schedule, that's worth $60M. What's your time-to-business value?

Ease of Use. While considering new Big Data solutions, it is imperative to consider how the solution will affect your need to augment internal resources for implementation and ongoing maintenance. Specifically, new resources, such as data scientists, may be necessary to handle a project internally. In addition, corporate IT resources will be needed to maintain the project once implemented. Ease of use also translates into ease of integration within your internal technical infrastructure. This can have a big impact on time-to-business value.

Scalability. Ideal solutions will expand with evolving business needs. While one use case may be the initial driver, the ideal solution will support future uses cases. Also, consider any costs associated with scaling. (Some solutions offer initial capacity at a low cost, but the cost quickly increases as expansion occurs.)

> **At Infochimps**, we've see customers enduring time to business value of 18-24 months with legacy solutions, and where time-to-value can be as short as 30 days with cloud-based deployments.

Standards-Based. In today's world, popular technologies come and go. Consider whether your proposed solutions use open standards-based, best-in-class technologies to drive innovation and revenue for your business needs. Additionally, explore any risks associated with being able to tailor your solution and your ability to act quickly when you need to do so.

Enterprise Readiness. An important aspect of selection is whether the solution can operate within an enterprise setting. Ideal solutions have high availability and disaster recovery in place. Additionally, they will support your business compliance and security needs, and meet your current company defined SLAs. To better understand how solutions operate in an enterprise environment, contact the solution provider's enterprise customer references.

3-year Total Cost of Ownership. You need to add up your costs over at least three years to appreciate some of the recurring costs which may not be completely obvious the first year. Costs include allocated data center infrastructure (floor, racks, power, connectivity), hardware (compute, storage, and networking), software (Big Data stack, OS, Admin SW, security SW, analytics SW), personnel costs (systems management/NOC, development, consulting).

Research shows that nearly 55% of Big Data projects fail. At Infochimps, we are experts in Big Data projects,implementing numerous enterprise projects that deliver insights within 30 days. This Big Data guideline is based on a culmination of our successes; we hope you find value in it. If you would like to hear more about the Infochimps Cloud, please *request a demo*. If you would like to discuss your project, *please request a complimentary consultation*.

Project Overview

Accountable Executive	Department	Impact

Objective

Expected Outcome

Approach

Phase 1 : Pilot

Phase 2 : Production

Success Measures

Pilot Success

Production Success

Estimated Cost

Category	Description	Initial Cost	3yr TCO
Personnel			
Software			
Hardware			
Training			
Consulting			
Time to Value			

Activity & Timing

Phase	Major Activities	Timing
Initiation & Planning		
Execution		
Closure		

Deliverables

Name	Description	Timing
Project Plan		
Infrastructure Functional Spec		
Implementation Spec		

Dependencies, Assumptions & Constraints

Technical Requirements

Solution Description

Data Inputs

Name	Example	Feed 1	Feed 2	Feed 3	Feed 4
Listener Type *HTTP Req./Stream, Syslog, Batch upload*	*HTTP Request*				
Data Type *JSON, XML, CSV/TSV/ Delimited, Fixed Width, SQL, Binary*	*JSON*				
# of Channels *1 or many; Gnip is 1 source with possibly 100s of diff. URLs to listen to*	*1*				
Avg. Events / Sec	*125*				
Max Sustained Events / Sec (10 min)	*500*				
12-Month Growth Factor of Data *How much volume is expected to grow in next year?*	*5x*				
Avg. Event Size	*0.4KB*				
Streaming Aggregation?	*By Minute*				
Non-Trivial Decorators *Calls to external APIs/ data stores, complex algorithms/logic, &c.*	*Sentiment Analyzer*				

Batch Jobs

Name	Example	Job 1	...	Job N
Description *What does this job do?*	*Take last week's raw data and bin by hour*			
Frequency *How often does it run?*	*Nightly*			
Input(s) *Where is data from?*	*S3*			
Output(s) *Where does data write to?*	*MySQL*			
# of Records *Avg. # of records in each run*	*1 B*			
Requires Persistent Data? *Do we have to store any data on the Hadoop cluster itself to perform the job?*	*No*			

Type	Wukong	Hive	Java M/R	Pig
Hadoop Development Method *What method should be used to perform Hadoop jobs?*				

Data Stores Examples

Type	ElasticSearch	HBase	MySQL	HDFS	S3/Glacier
Purpose	*Support a customer-facing web app*	*Support an internal BI tool*	*Support a Tableau installation*	*Allow ad hoc queries in Hive*	*Archival/Disaster Recovery*
Avg. Events / Sec	*50*	*200*	*10*	*200*	*200*
Max. Sustained Events / Sec (10 min)	*200*	*500*	*50*	*500*	*500*
Retention Policy	*Keep all events from last 30 days*	*Keep all events from last 90 days*	*Keep all events from last 7 days*	*Keep all events from last 6 months*	*Keep all events from last 12 months.* *Last 3 months in Glacier*
12-month Growth Rate	*3x*	*3x*	*3x*	*3x*	*3x*
Query Profile *What kinds of queries/ usage/volume should be supported?*	*~10 simple searches / sec.* *Fewer complex/ faceted queries*	*~20 key/value lookups / sec.* *Many short table scans to create time series*	*~5 joins and time series queries / sec to support Tableau*	*Ad hoc queries running over large amounts of historical data*	*Occasional batch jobs from Hadoop*

Additional Tools

Any Additional Tools?	

Big Data Systems & Deployment Enviroment

Type	Public Cloud	Virtual Private Cloud	Private Cloud	Hybrid Cloud
Enviroment				

	ETL + Stream Processing	Databases	Hadoop
Production Nodes	*4*	*4 (ElasticSearch)*	*1 Elastic Cluster*
Staging Nodes	*2*	*2 (ElasticSearch)*	*n/a*

SLA Requirements

	Example	Production	Staging
System Uptime	*99.99%*		
Query Latency Response Time	*<200ms*		
Data Retention	*99.99%*		

Solution Diagram

Request a Free Demo

See Infochimps Cloud for Big Data.

Infochimps Cloud is a suite of enterprise-ready cloud services that make it simpler, faster and far less complicated to develop and deploy Big Data applications in public, virtual private and private clouds.

Our cloud services provide a comprehensive analytics platform, including data streaming, storage, queries and administration. With Infochimps, you focus on the analytics that drive your business insights, not building and managing a complex infrastructure.

INFOCHIMPS.COM/DEMO ➜

Contact Us

Infochimps, Inc
1214 W 6th St. Suite 120
Austin, TX 78703
1-855-328-2386
www.infochimps.com
info@infochimps.com
Twitter: @infochimps