Whitepaper

The guide to successfully building Big Data solutions

The guide to successfully building Big Data solutions

Big Data analytics is an indispensable tool for supporting the efficient business decision-making. There are multiple cloud-based solutions for Big Data analytics and data visualization tools, like Jupyter, Tableau, Google Chart or D3.js. Tableau is actually amongst the leaders in Gartner's BI and Analytics Platforms as of 2017.

Figure 1. Magic Quadrant for Analytics and Business Intelligence Platforms

Source: Gartner (February 2018)

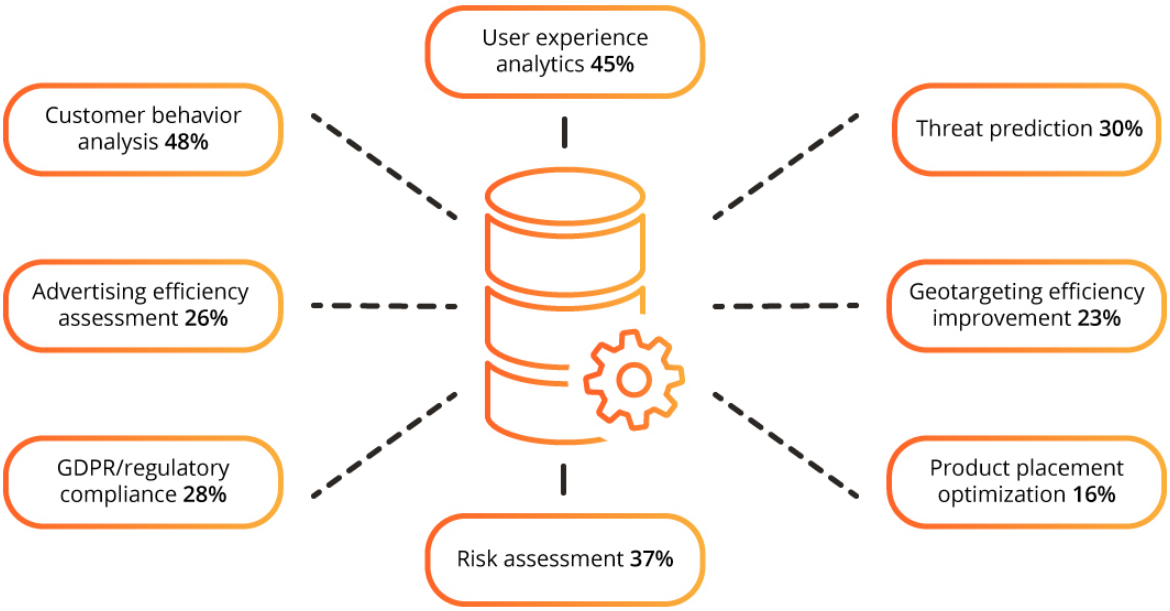
However, many data visualization systems cannot run with legacy systems, thus the need for bespoke Big Data analytics solutions and customization of the existing platforms and tools. IT Svit has extensive experience with this task, and this white paper is our guide to successfully building Big Data solutions.

1 Step

Task analysis

Building Big Data solutions for information processing and analytics is the task IT Svit is frequently asked to accomplish. First of all, here are some stats on such projects, based on our experience and the data from public sources:

1. Possible applications of Big Data projects

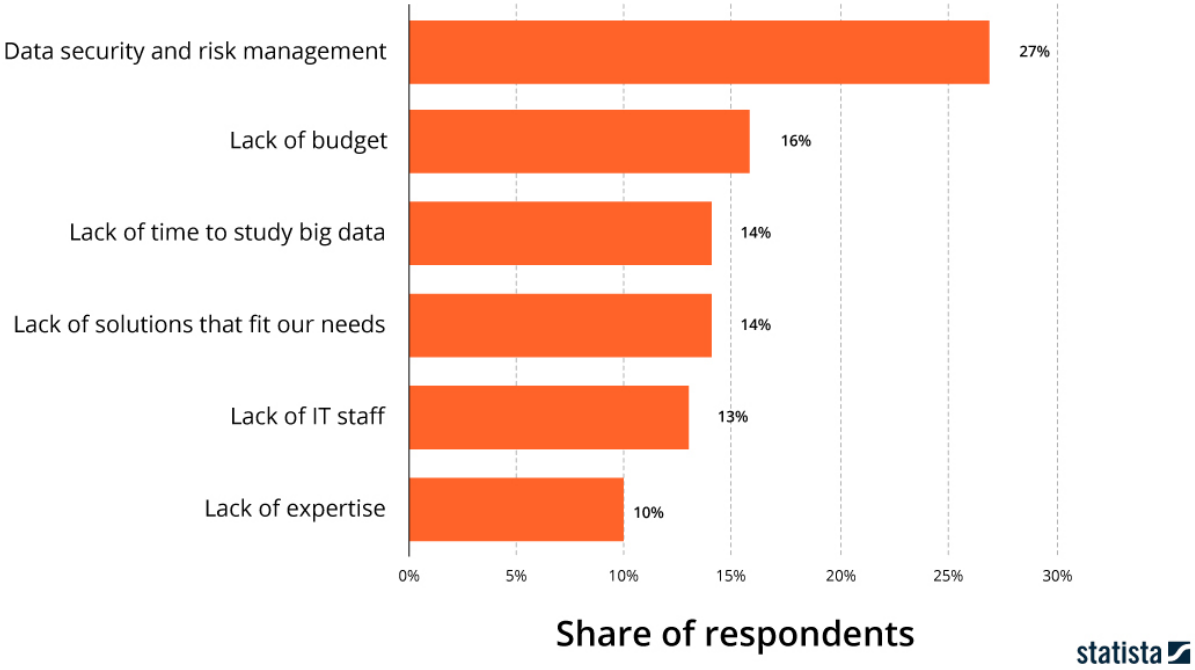


2. Approximate breakdown of costs, risks and time expenses

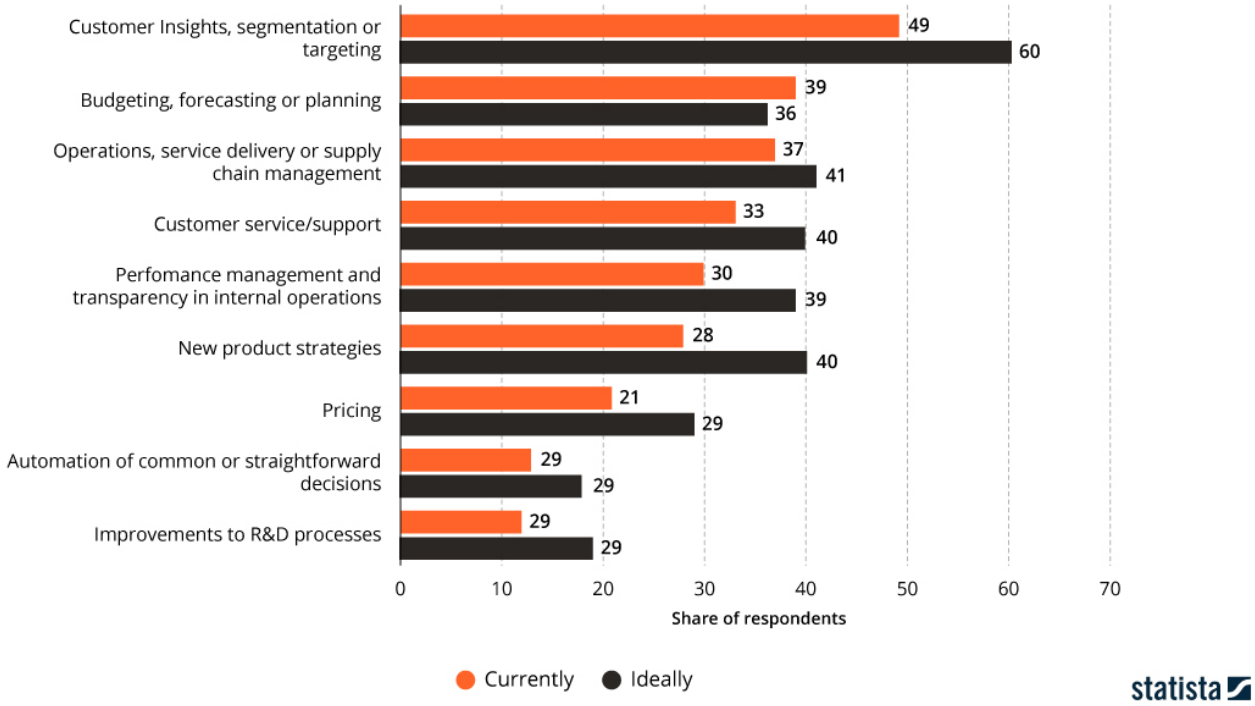
Expenses / Stage	Cost, %	Time, %	Risk, %
Task analysis	10	10	40
Solution design	10	30	30
Solution implementation	50	40	25
Solution refinement and scaling	30	20	5

As you can see, the task analysis stage is usually relatively short and costs little, yet it is the most risky stage. Many projects never go past the risk assessment, which is surely for the best.

3. The biggest challenges when working with Big Data, according to Statista:



4. The areas where Big Data analytics can increase the business efficiency.



This chart is based on the responses of more than 1500 SEOs and business executives who have currently implemented or plan to implement Big Data analytics solutions.

2 Step

Big Data Solution design

The main priority of the pilot build of the project is to test the Proof of Concept — if the system is able to work at all. Thus said, the technology stack chosen should ensure the simplicity and speed of development, as well as appropriate performance of the finished solution. The solution might have to be redesigned from scratch if an impassable block is met during the implementation phase. That is why the design stage is the second most risky and the second most expensive of all four, despite still being relatively short.

To design a good Big Data analytics solution one needs a decent understanding of Big Data architecture. The type of input data dictates the required technology stack:

- Processing textual data requires using non-relational databases like Cassandra or MongoDB, writing scripts in Python with async.io for parallel processing, and working with Hadoop set of utilities and Spark framework.
- Processing video and images requires using OpenCV with geodata tools and smart triggers
- Processing audio files requires the tools for speech-to-text recognition.

The ultimate goal of the data processing is transforming it to enable storing the JSON into the database.

Once the customer requests us to develop a Big Data analytics solution for them, we have to answer the following questions:

1. Determining the list of data sources, their validity and priority.

1 What is the list of the sources of data we will have to process? What is the valuable, high-priority data we need?

Data Source	Total % Asked to Analyze
Documents	84%
Business transactions in database	82%
E-mail	74%
Imaging data	68%
Sensor or device data	57%
Internet search indexing	57%
Weblogs	55%
Social media	54%
Telephone conversations	52%
Videos	52%
Pictures	46%
Clickstreams	42%
Other	2%

2. Determining the amount of computing resources needed.

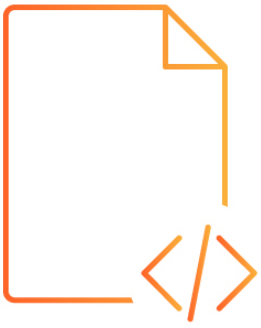
2 Is such process realistic at all? What amount of computing resources will be needed?

3. Determining the technological stack needed to implement the solution.

3 What technologies will the solution use?

4. Determining the data relevancy and value criteria.

4 How to ensure the data is up-to-date, relevant and meets the value criteria ?



Python scripts

5. **Determining the ways to remove the duplicates and find the original sources of data.**

5 How to remove null-weight data, and duplicates?
How to find the original sources of data?

Google Search APIs

6. **Determining the needed team composition and preferred project management model.**

6 What team and time scope is needed to accomplish the project? Will the project be feasible?

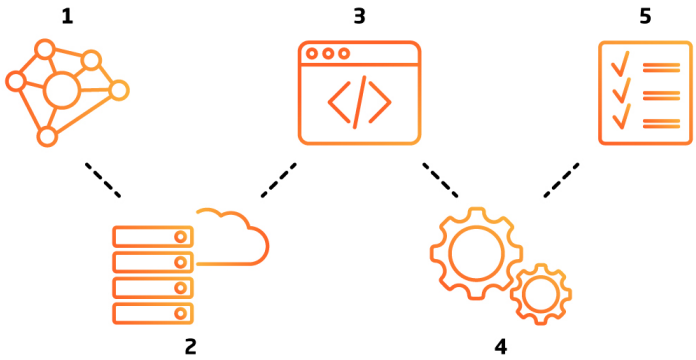
3 Step

Solution implementation

This stage might account for up to 50% of the whole project time, costs and risks. During this period the team might meet an unavoidable roadblock in development and come to a conclusion a complete reevaluation of the concept or a solution redesign is required. This is the sad reality of Big Data app development, yet this stage will result in the first working MVP and provide the first tangible results.

The tasks that we need completed here:

1. Provisioning the infrastructure, setup of automatic backups and restoration procedures
2. Provisioning and configuring the databases and their replication
3. Developing, testing and deployment of solution code
4. Project management of the software delivery pipeline
5. QA of the solution code, checking its integrity, compatibility and security



Once the solution begins providing the data we must begin analyzing the credibility and usefulness of the received results. **Do they solve the task outlined on the first stage?** Are they credible and can they support the decision-making or carry out any other function implied?

If the results do not meet the requirements of the task, a reassessment of the technology and data processing used might lead to new iteration of the solution design. These pilot run results must answer one more question: is the solution production-ready and does it scale well?

4 Step

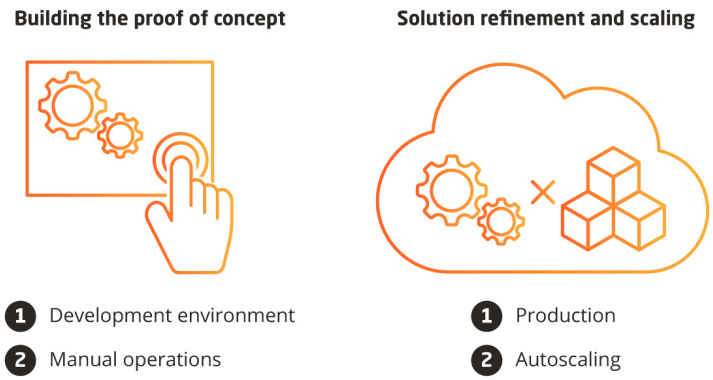
Solution refinement and scaling

As the design and implementation stages served merely the purpose to deliver some positive results as quickly and easily as possible, the refinement and scaling stage must end with the finished and polished product that can be scaled up and down as needed. The risk is minimal on this stage, yet it still demands a significant amount of time and resources.

This can be ensured by the following procedures:

- Code refactoring
- Feature adjustment
- Redesigning for scaling
- Cloud migration
- Integration with decision-making support systems
- Providing easily-discernible reporting and data visualization for business analytics
- Ongoing technical support and evolution of the solution

It might turn out the solution must be rebuilt from scratch leveraging the cloud-based features to meet the scaling demands. Building with these in mind from the very beginning would be the best course of actions, yet more often than not the final product works somewhat differently than was expected from the beginning. The initial development might serve only as the way to gain a clear understanding of how the solution parts must work and interact with each other.



The only way to avoid the need to rebuild the solution from scratch and meet the initial timeframes and budget estimates is by working with experienced contractors that have ample experience with building efficient Big Data solutions.

About IT Svit

We help industry-leading companies in the US, EU and worldwide innovate and overcome their challenges. IT Svit does this using enterprise-grade technology solutions to drive value and ensure business continuity. We pride ourselves on delivering great user experience and brand advocacy.

IT Svit specializes in DevOps services, Big Data technology, Machine Learning, bespoke Blockchain platforms, full-cycle services for startups, web development and QA.