ScienceSoft
Professional Software Development

www.scnsoft.com

Big Data Implementation

A Practical Guide

US, McKinney, TX +1 214 306 68 37 Finland, Vantaa +358 92 316 30 70 UAE, Fujairah +971 585 73 84 33
www.scnsoft.com contact@scnsoft.com

Contributors

Alex Bekker

Head of Data Analytics Department, ScienceSoft

Marina Chernik

PhD, Senior Business Analyst and BI Consultant, ScienceSoft

Big Data Solution Implementation: The Essence

Big data implementation gains strategic importance across all major industry sectors, helping mid-sized and large organizations successfully handle the ever-growing amount of data for operational and analytical needs. When properly developed and maintained, big data solutions efficiently accommodate and process petabytes of XaaS users' data, enable IoT-driven automation, facilitate advanced analytics to boost enterprise decision making, and more.

Required team: Project manager, business analyst, big data architect, big data developer, data engineer, data scientist, data analyst, DataOps engineer, DevOps engineer, QA engineer, test engineers. Check typical roles on ScienceSoft's big data teams.

Costs: from **$200,000** to **$3,000,000** for a mid-sized organization, depending on the project scope. Use our free calculator to estimate the cost for your case.

For 10 years, ScienceSoft designs and builds efficient and resilient big data solutions with scalable architectures able to withstand extreme concurrency, request rates, and traffic spikes.

Key Components of a Big Data Solution

Below, ScienceSoft's big data experts provide an example of a high-level big data architecture and describe its key components.

- **Data sources** provide real-time data (e.g., from payment processing systems, IoT sensors) and historical data (e.g., from relational databases, web server log files, etc.).
- **Data storage (a data lake)** holds multi-source voluminous data in its initial format (structured, unstructured, and semi-structured) for further batch processing.
- A stream ingestion engine captures real-time messages and directs them to the real-time (stream) processing module and the data lake (the latter stores this data to ensure its in-depth analytics during further batch processing).

Batch processing engine

Best for: repetitive non-time-sensitive jobs that facilitate analytics tasks (e.g., billing, revenue reports, daily price optimization, demand forecasting).

- Deals with large datasets.
- The results are available as per the established computation schedule (e.g., every 24 hours).

Stream processing engine

Best for: tasks that require an immediate response (e.g., payment processing, traffic control, personalized recommendations on ecommerce websites, burglary protection systems).

- Deals with small data chunks.
- The processed data is always up-to-date and ready for immediate use due to low latency (milliseconds to seconds).

Note: *Depending on your big data needs, a specific solution might enable only batch or only stream data processing, or combine both types as shown in the sample architecture above.*

- Once processed, data can go to a **data warehouse** for further analytical querying or directly to the analytics modules.
- Lastly, the **analytics and reporting** module helps reveal patterns and trends in the processed data, then use these findings to enhance decision-making or automate certain complex processes (e.g., management of smart cities).
- With **orchestration** that acts as a centralized control to data management processes, repeated data processing operations get automated.

Big Data Implementation Steps

Real-life big data implementation steps may vary greatly depending on the business goals a solution is to meet, data processing specifics (e.g., real-time, batch processing, both), etc. However, from ScienceSoft's experience, there are six universal steps that are likely to be present in most projects.

Step 1. Feasibility study

Analyzing business specifics and needs, validating the feasibility of a big data solution, calculating the estimated cost and ROI for the implementation project, assessing the operating costs.

> Big data initiatives require a thorough feasibility investigation to avoid unnecessary expenses. To ensure that each dollar spent brings our customers real value, ScienceSoft's big data consultants prepare a comprehensive feasibility report featuring tangible gains and possible risks.

Marina Chernik
PhD, Senior Business Analyst and BI Consultant, ScienceSoft

Step 2. Requirements engineering and big data solution planning

- Defining the types of data (e.g., SaaS data, SCM records, operational data, images and video) to be collected and stored, the estimated data volume, and the required data quality metrics.
- Forming a high-level vision of the future big data solution, outlining:
  - Data processing specifics (batch, real-time, or both).
  - Required storage capabilities (data availability, data retention period, etc.).
  - Integrations with the existing IT infrastructure components (if applicable).
  - The number of potential users.
  - Security and compliance (e.g., HIPAA, PCI DSS, GDPR) requirements.
  - Analytics processes to be introduced to the solution (e.g., data mining, ML-powered predictive analytics).
- Choosing a deployment model: on-premises vs. cloud (public, private) vs. hybrid.
- Selecting an optimal technology stack.

- Preparing a comprehensive project plan with timeframes, required talents, and budget outlined.

ScienceSoft can provide you with expert guidance on all aspects of big data planning.

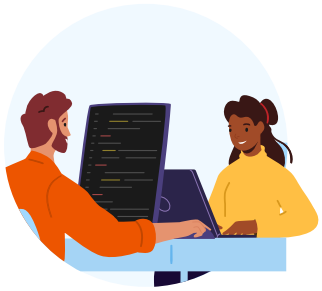
Step 3. Architecture design

- Creating the data models that represent all data objects to be stored in big data databases, as well as associations between them, to get a clear picture of data flows, the ways data of certain formats will be collected, stored, and processed in the solution-to-be.
- Mapping out data quality management strategy and data security mechanisms (data encryption, user access control, redundancy, etc.).
- Designing the optimal big data architecture that enables data ingestion, processing, storage, and analytics.

> As your business grows, the number of big data sources and the overall data volume will likely grow too. For instance, if we compare infographics for 2020 and 2021, we'll see that the volume of video streams on YouTube per minute grew from 500 to almost 700 hours in just a year. This makes scalable architecture the cornerstone of efficient big data implementation that can save you from costly redevelopments down the road.

Alex Bekker
Head of Data Analytics Department, ScienceSoft

Step 4. Big data solution development and testing



- Setting up the environments for development and delivery automation (CI/CD pipelines, container orchestration, etc.).
- Building the required big data components (e.g., ETL pipelines, a data lake, a DWH) or the entire solution using the selected techs.
- Implementing data security measures.
- Performing quality assurance in parallel with development. Conducting comprehensive testing of the big data solution,

including functional, performance, security and compliance testing. If you're interested in the specifics of big data testing process, see expert guide by ScienceSoft.

With 10 years of experience in delivering end-to-end big data solutions, ScienceSoft is ready to help you develop and deploy your big data software.

Step 5. Big data solution deployment

- Preparing the target computing environment and moving the big data solution to production.
- Setting up the required security controls (audit logs, intrusion prevention system, etc.).
- Launching data ingestion from the data sources, verifying the data quality (consistency, accuracy, completeness, etc.) within the deployed solution.
- Running system testing to validate that the entire big data solution works as expected in the target IT infrastructure.
- Selecting and configuring big data solution monitoring tools, setting alerts for the issues that require immediate attention (e.g., server failures, data inconsistencies, overloaded message queue).
- Delivering user training materials (FAQs, user manuals, a knowledge base) and conducting Q&A sessions and trainings, if needed.

Step 6. Support and evolution (continuous)

- Establishing support and maintenance procedures to ensure trouble-free operation of the big data solution: resolving user issues, refining the software and network settings, optimizing computing and storage resources utilization, etc.
- Evolution may include developing new software modules and integrations, adding new data sources, expanding the big data analytics capabilities, introducing new security measures, etc.

Sourcing Models for Big Data Solution Implementation

In-house big data solution development

+ Full control over the project.

− Possible lack of talents with required skills, insufficient team scalability.
− All managerial efforts are on your side.

Team augmentation

+ On-demand availability of talents with the required skills.
+ Sufficient control over the project.

− Extra efforts to achieve smooth cooperation between the teams.

Outsourced big data solution development

+ A fully managed & experienced big data team.
+ Established best practices for big data implementation.
+ A quick project start.

− Risks of choosing an inept vendor.

Selected Big Data Projects by ScienceSoft

Development of a Big Data Solution for IoT Pet Trackers

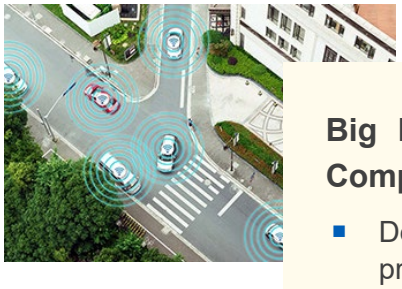
- Design and development of an easily scalable big data solution that processes 30,000+ events per second from 1 million devices.
- Enabling real-time pet location tracking, as well as sending and receiving photos, videos, and voice messages via an app.
- Setting automatic hourly, weekly, or monthly reports with the option to tune the reporting period.

PROJECT DETAILS →

Big Data Implementation for Advertising Channel Analysis

- Development of a new analytical system that handles the continuously growing amount of data and enables advertising channel analysis in 10+ countries.
- Processing more than 1,000 different types of raw data (archives, XLS, TXT, etc.).
- Enabling cross analysis of almost 30,000 attributes and facilitating multi-angled data analytics for different markets.

PROJECT DETAILS →



Big Data Consulting for a Leading Internet of Vehicles Company

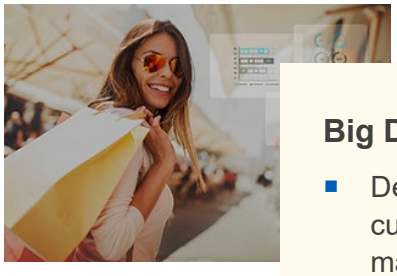
- Design and development of an easily scalable big data solution that processes 30,000+ events per second from 1 million devices.
- Enabling real-time pet location tracking, as well as sending and receiving photos, videos, and voice messages via an app.
- Setting automatic hourly, weekly, or monthly reports with the option to tune the reporting period.

PROJECT DETAILS →

Big Data Consulting and Training for a Satellite Agency

- Preparing comprehensive educational materials to introduce the client to the big data landscape with a focus on the space industry.
- Training sessions to the top management and technical team in the form of workshops with Q&A sessions.
- In-depth analysis of strong and weak points of the planned big data solution's architecture.

PROJECT DETAILS →



Big Data Implementation for a Multibusiness Corporation

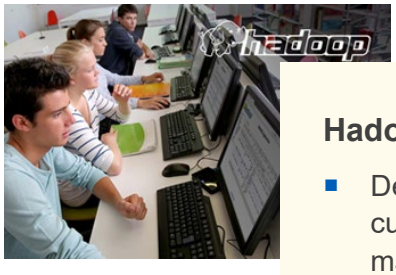
- Development of a big data solution that offered a 360-degree customer view as well as functionality for retail analytics, stock management optimization, and employee performance assessment.
- Setting up a data warehouse and around 100 ETL processes.
- An analytical server with 5 OLAP-cubes and about 60 dimensions in total.

PROJECT DETAILS →

Apache NiFi Managed Support for a Biotechnology Corporation with 10,000+ Employees

- Preparing comprehensive educational materials to introduce the client to the big data landscape with a focus on the space industry.
- Training sessions to the top management and technical team in the form of workshops with Q&A sessions.
- In-depth analysis of strong and weak points of the planned big data solution's architecture.

PROJECT DETAILS →



Hadoop Lab Deployment and Support

- Development of a big data solution that offered a 360-degree customer view as well as functionality for retail analytics, stock management optimization, and employee performance assessment.
- Setting up a data warehouse and around 100 ETL processes.
- An analytical server with 5 OLAP-cubes and about 60 dimensions in total.

PROJECT DETAILS →

The Benefits of Implementing Big Data with ScienceSoft

Improved business processes

Having practical experience in 30+ industries, we know how to create solutions that not only fit our customers' unique processes but improve them.

In-depth data analytics

In data analytics and AI/ML since 1989, we know which techs and approaches to apply to gather disparate data into a single point of truth and ensure timely and accurate analytics.

Top-notch UX

We build easy-to-navigate interfaces, create informative visuals, and implement self-service capabilities to make data exploration and presentation a smooth experience, even for non-technical users.

Complete security

With our ISO 27001-certified security management system and zero security breaches in the company's entire 34-year history, we can guarantee full protection of your big data software.

Focus on future-proofing

We build big data systems that preserve stable performance from day one and can be easily scaled and upgraded to accommodate any data volume increase or ensure new capabilities in the future.

Vendor neutrality

We are proficient in 50+ big data techs and hold official partnerships with Microsoft, AWS, and Oracle. So, when we choose the techs, we are guided only by the value they will drive in every particular project.

Why Choose ScienceSoft for Big Data Implementation

- 11 years in big data solutions development.
- 35 years in data analytics and data science.
- Experience in 30+ industries, including manufacturing, retail, healthcare, education, logistics, banking, energy, telecoms, and more.
- 750+ experts on board, including big data solution architects, DataOps engineers, and ISTQB-certified QA engineers.
- A Microsoft partner since 2008.
- An AWS Select Tier Services Partner.
- Strong Agile and DevOps culture.
- ISO 9001 and ISO 27001-certified to ensure robust quality management system and the security of the customers' data.
- ScienceSoft is a 3-Year Champion in The Americas' Fastest-Growing Companies Rating by the Financial Times.

Our Customers Say

Invention Machine

Mark Atkins
CEO

ScienceSoft has delivered cutting-edge solutions to complex problems bringing in innovative ideas and developments.

ScienceSoft follows specifications very rigidly, requiring clear communication about intended functionality. My final comment about ScienceSoft reflects their dedication to handle any problem that occurs as a result of hardware or software issues; simply put, they will go the extra mile to support their customers regardless of the time of day these issues arise.

Miami Dade College

Kaiyang Liang Ph.D
Professor

We needed a proficient big data consultancy to deploy a Hadoop lab for us and to support us on the way to its successful and fast adoption.

ScienceSoft's team proved their mastery in a vast range of big data technologies we required: Hadoop Distributed File System, Hadoop MapReduce, Apache Hive, Apache Ambari, Apache Oozie, Apache Spark, Apache ZooKeeper are just a couple of names. Whenever a question arose, we got it answered almost instantly.

Check the project

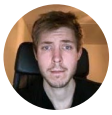
travision

Simen Løkka
CEO

We commissioned ScienceSoft to audit and upgrade our partially developed AI-based software for clay pigeon shooting tracking.

As a result, the system could track a flying target in a real-life outdoor environment and faultlessly detect shooter’s performance. We are satisfied with our cooperation with ScienceSoft and their skilled development team, which smoothly fit into our project. In case of further system evolution, we’ll continue our collaboration and won’t hesitate to recommend ScienceSoft as a reliable development partner.

YAGER



Joakim Ohlander
Technical Director

ScienceSoft has been a life savior for us and our players when we were about to release our video game The Cycle Frontier and were facing immediate issues in terms of backend scalability.

Their combination of expert knowledge at Microsoft Azure .NET and great agile collaboration skills allowed us to start working fast and effectively together in solving problems which allowed us to release.

Check the project

Typical Roles on ScienceSoft's Big Data Teams

With 750+ professionals on board, ScienceSoft has experts to cover any project scope. Below, we describe key roles in our big data projects teams that can also be augmented with our front-end developers, UX and UI designers, BI engineers, and other professionals.

Project manager

Plans and oversees a big data implementation project; ensures compliance with the timeframes and budget; reports to the stakeholders.

Business analyst

Analyzes the business needs or app vision; elicits functional and non-functional requirements; verifies the project's feasibility.

Big data architect

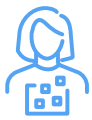
Works out several architectural concepts to discuss them with the project stakeholders; creates data models; designs the chosen big data architecture and its integration points (if needed); selects the tech stack.

Big data developer

Assists in selecting techs; develops big data solution components; integrates the components with the required systems; fixes code issues and other defects on a QA team's notices.

Data engineer

Assists in creating data models; designs, builds, and manages data pipelines; develops and implements a data quality management strategy.



Data scientist

Designs the processes of data mining; designs ML models; introduces ML capabilities into the big data solution; establishes predictive and prescriptive analytics.

Data analyst

Assists a data engineer in working out a data quality management strategy; selects analytics and reporting tools.

DataOps engineer

Helps streamline big data solution implementation by applying DevOps practices to the big data pipelines and workflows.

DevOps engineer

Sets up the big data solution development infrastructure; introduces CI/CD pipelines to automate development and release; deploys the solution into the production environment; monitors solution performance, security, etc.

QA engineer

Designs and implements a quality assurance strategy for a big data solution and high-level testing plans for its components.

Test engineer

Designs and develops manual and automated test cases to comprehensively test the operational and analytical parts of the big data solution; reports on the discovered issues found and validates the fixed defects.

Technologies ScienceSoft Uses to Develop Big Data Solutions

Distributed storage	Hadoop, Amazon Simple Storage Service, Azure Blob Storage
Database management	Apache Cassandra, Azure Cosmos DB, Azure Synapse Analytics, Amazon DynamoDB, Amazon Redshift, Amazon DocumentDB, Apache Hive, Apache HBase, Apache NiFi, Google Cloud Datastore, MongoDB
Data management	Apache Airflow, talend, Informatica, ZALONI, Apache ZooKeeper, Azkaban
Data streaming and stream processing	kafka, Apache NiFi, IoT Hub, amazon Kinesis, Apache Spark, Apache Storm, Azure Stream Analytics
Batch processing	Hadoop MapReduce, amazon EMR, Apache Hive, Apache Pig
Data warehouse, ad hoc exploration and reporting	PostgreSQL, Azure Synapse Analytics, amazon REDSHIFT, Power BI, tableau, QlikView, Google Developers Charts, Grafana, SISENSE
Machine learning	Matlab, Octave, R, MAHOUT, Caffe, mxnet, TensorFlow, Keras, torch, OpenCV, theano, Apache Spark ML, Apache Spark MLlib, scikit-learn, GENSIM, spaCy, amazon Machine Learning, Amazon SageMaker, Azure Machine Learning, Google Cloud AI Platform, einstein
Programming languages	Scala, Java, python, R, C++

About ScienceSoft

ScienceSoft is a global IT consulting and software development company headquartered in McKinney, TX. Since 2013, we have been delivering end-to-end big data services to businesses in 30+ industries. Being ISO 9001 and ISO 27001-certified, we ensure robust quality management system and full security of our customers' data.