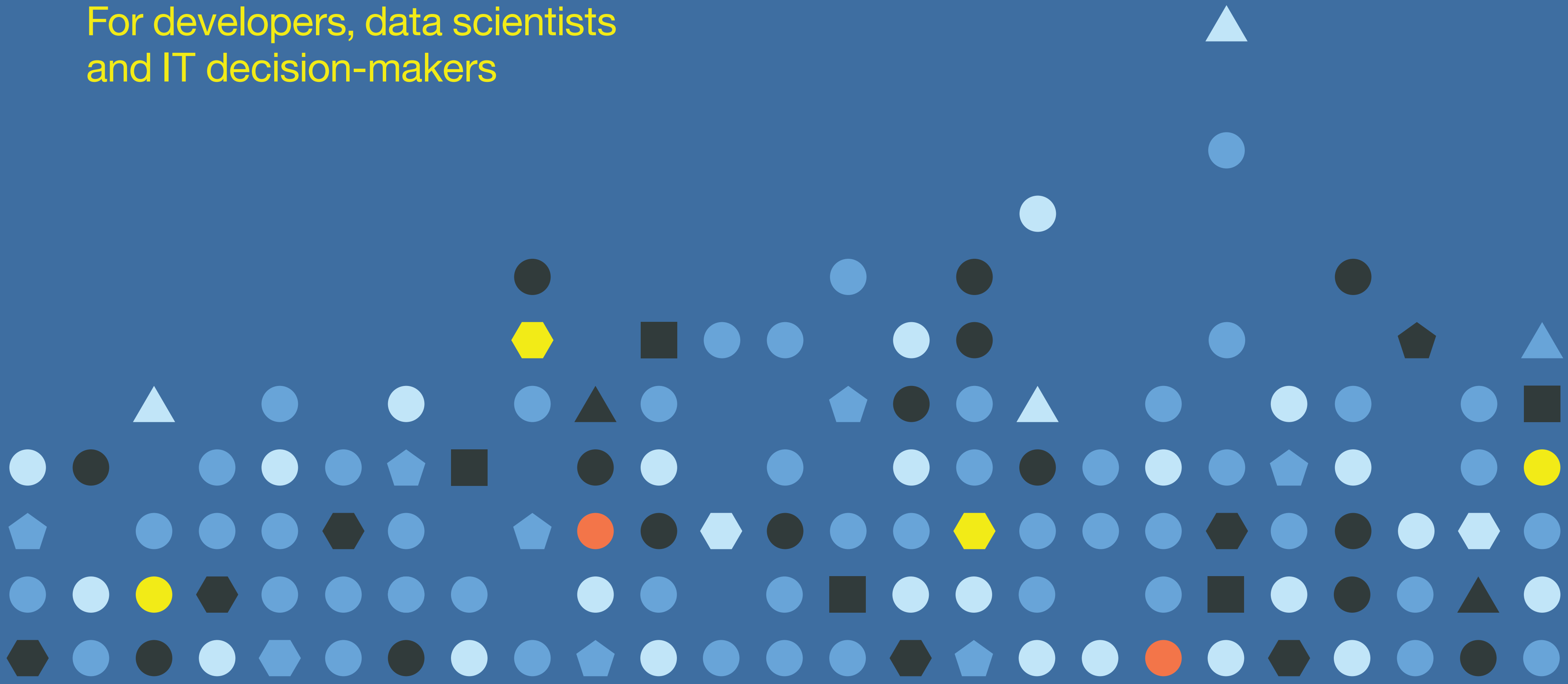
Cloud Data Services

Getting Faster Insights from Application Data with IBM Cloudant and Apache Spark

For developers, data scientists and IT decision-makers



Highlights

NoSQL databases and Apache Spark are a potent combination for rapid integration, transformation and analysis of all kinds of business data.

With its data syncing and analytics capabilities, IBM Cloudant offers unique advantages as a NoSQL database for many Spark use cases.

IT decision-makers, data scientists and developers need to know how and when to apply these technologies most effectively.

IBM can offer a host of resources and tools to help your organization gain value from Cloudant and Spark quickly, and with minimal up-front investment.

1 Introduction

The emergence of Apache® Spark™ marks the beginning of a new era in analytics. Unlike earlier big data technologies such as Hadoop MapReduce, Spark's data processing engine runs entirely in-memory—which enables it to integrate, transform and analyze truly large datasets and semi-structured data 10 or even 100 times faster than ever before.[1]

What does this mean in practical terms? Apache Spark opens up a new world of possibilities for forward-thinking organizations to explore, understand and extract value from data. Most organizations have sources of data that are considered too difficult to analyze, or cannot be analyzed within a reasonable time-frame. Spark makes it possible to shine a light through this previously opaque data, and bring new insights into sharp focus—often within seconds.

Much of the data that Spark is designed to process is captured, stored and managed using NoSQL databases, which have gained increasing popularity over the last decade. NoSQL databases like IBM® Cloudant®, a JSON document store, provide a flexible and reliable way to persist large data sets at web-scale. They can also eliminate much of the traditional complexity and cost of data management.

Moreover, for developers and data scientists in both startups and more established companies, the ease-of-use and flexible schemas of NoSQL databases empower small teams or individuals to build applications and develop algorithms quickly, simply, and with minimal risk.

IBM makes it even easier to take advantage of both Spark and NoSQL technologies by offering them together, as flexible, affordable cloud data services. Spinning up a new Cloudant database and integrating it with IBM Analytics for Apache Spark is the work of a few minutes, thanks to IBM's open source Spark-Cloudant Connector and the IBM Bluemix® cloud development platform.

Three key groups—IT decision-makers, data scientists and developers—have the ability to harness Spark and Cloudant, and turn previously untapped datasets into new sources of business value.

2 For the IT decision-maker:

Why are Cloudant and Spark important for my business?

The rapid rise of tech-enabled startups has shown that converting an innovative idea into a successful business no longer requires a multimillion-dollar budget, a multi-year timeline, a stack of expensive infrastructure and a large project team.

There is a simple reason why innovation is now less expensive and risky than ever before: software development practices have evolved to a point where projects can be started and delivered quickly, with minimal up-front costs. This makes it easier for even the most risk-averse companies to say “yes” when an opportunity arises to enter a new market, create a new line of business, or disrupt their competitors.

As a result, the business landscape has changed forever. Small companies are suddenly finding it much easier to compete against larger rivals. Meanwhile, large companies are realizing that they need to start thinking like smaller, more agile businesses in order to defend their market share. In this new business climate, the role of IT decision-makers is to identify opportunities to innovate—that their organizations have the technical resources to seize these opportunities when they arise.

At the moment, there are no more fertile frontiers for innovation than the emerging realms of cognitive computing, big data analytics, and the Internet of Things. This paper aims to explain why Cloudant and Spark are the right technologies to help you solve the key challenges in these new domains.

Before we take a deeper dive, let’s take a look at an example of how one company is already benefiting from integration between Cloudant and Spark.

Case study:

Smarter maintenance with the Internet of Things

Any company that needs to manage thousands of complex physical assets understands that maintenance planning can present a significant challenge. Knowing when and where an asset is likely to fail represents a key competitive advantage, because it makes it possible to plan maintenance cycles more efficiently and minimize costly downtime.

One IBM client, an engineering and service company, realized that it could use the Internet of Things to solve this problem. It wanted to use Internet-connected sensors to continuously send statistics on the movements, loads and failures of the equipment that it manages, combine that data with information captured by engineers during maintenance visits, and then analyze all the data to find opportunities for smarter maintenance. However, the company supports many different brands and models of equipment, which all have their own different sensors and data formats.

As a result, it was difficult to design a relational database schema that would be capable of storing all the different types of data that needed to be analyzed.

Relational database schemas also tend to be inflexible, which raised additional potential problems for the company. For example, if information from a new type of sensor needed to be added to the dataset in the future, it would have been difficult to adapt the existing schema to the new requirements.

IBM Cloudant and Spark provided the answer. Instead of requiring the company to design a rigid schema at the start of the project, Cloudant allowed it to ingest all of the information and store it as a series of JSON documents. As a result, the data from each type of sensor can be stored as-is, without needing to undergo a complex transformation and normalization process.

The company then used Spark to analyze the data. Unlike traditional analytics tools, Spark does not need to make hard assumptions about the structure of the data it analyzes. It can automatically infer the schema of the data at runtime, and can even deal with datasets where the format varies from record to record.

This combination of technologies has enabled the company to get the new analytics platform up and running quickly, delivering insights that make it possible to make smarter decisions about asset maintenance, and potentially reducing downtime and costs.

The bottom line: *Cloudant and Spark enabled the rapid development of the new analytics platform, slashing the cost, complexity and risk of delivering this valuable capability to the business, and enabling more flexibility for the future.*

3 For the data scientist:

How can Cloudant and Spark help me explore and analyze my data?

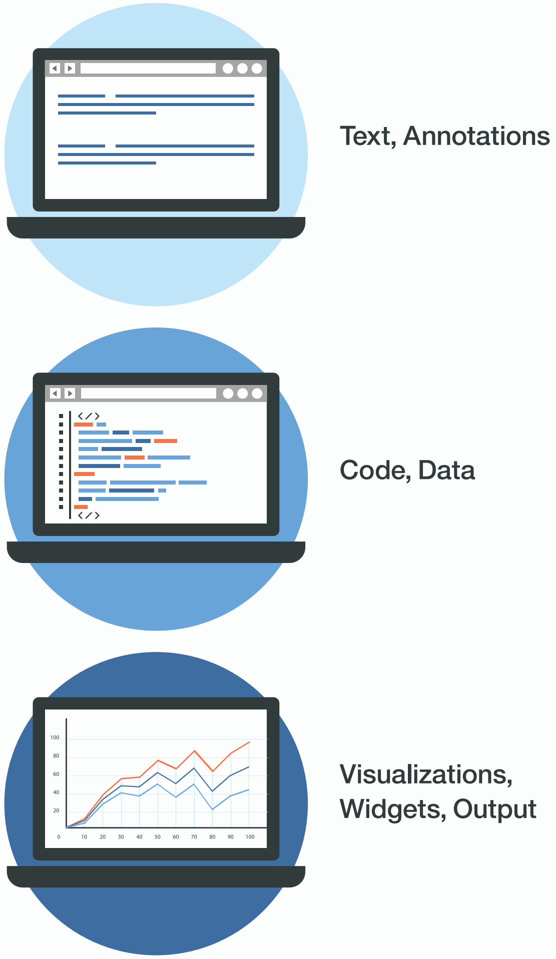
In the past few years, data science has emerged from its roots as an academic discipline to become one of the hottest professions in the business world. As organizations realize that their future depends on the value they can extract from their data, people who have the skills to explore and analyze that data can command a premium.

The rise of data science as a business discipline has also been facilitated by the availability of tools that make it easier to visualize data and communicate it to a wider business audience. Jupyter Notebooks, for example, provide a simple yet powerful environment both for the exploration and analysis of data (algorithms written in Python, R or Scala code), and for the generation of charts, graphics and narrative to describe the methodology and explain the results.

To complement Notebooks as a powerful user interface for data science, Cloudant offers a powerful back-end for data management. Its fully managed JSON data layer makes it easy for data scientists to store and manage datasets of any size, without needing the skills of a traditional database administrator to design complex schemas and table structures.

Each record within the Cloudant database exists as a self-contained “document” in JSON format, with no complex dependencies on other records. So it’s easy to load, explore and analyze anything from a single record to an entire database. Effectively, if you understand JSON data structures, you can use Cloudant (see **https://docs.cloudant.com/try.html** for an interactive demo).

What is a Notebook?



For the data scientist:

How can Cloudant and Spark help me explore and analyze my data?

Moreover, Cloudant is available as a cloud-based service, fully managed by IBM. This means that there is no need to spend time configuring servers or installing software—you can simply request a new Cloudant instance, load your data into it, and start working. And by using Cloudant in tandem with analytics services such as IBM Data Science Experience, you can connect a Notebook to a Cloudant data source with just a few mouse clicks.

Another shortcoming of the way most data scientists currently work with Notebooks is that big datasets also require big processing power—particularly for more sophisticated types of analysis such as machine learning. However, the standard editions of R and Python cannot easily take advantage of multi-core or multi-processor architectures—limiting the amount of heavy lifting that can be achieved on a single PC or even a powerful server.

Spark provides a solution to this problem by distributing the data across a cluster of servers, running the algorithm on each node in parallel, then aggregating the results and returning them to the user. And because all the calculations are performed on data held in memory, rather than on disk, Spark can complete even complex jobs on huge datasets very quickly. In many cases, long-running queries can be accelerated from hours down to minutes or even seconds.

The combination of Cloudant and Spark with Notebooks offers data scientists an extremely powerful environment for analyzing all kinds of big data. IBM's Spark-Cloudant Connector provides a pipeline to move data between the two environments seamlessly, and its ability to detect JSON schemas automatically means that no manual data preparation or transformation is required.

As businesses begin to focus more of their energies on mining sources of unstructured data from social networks and the Internet of Things, having the right tools in place for large-scale data science will be a key differentiator. Cloudant and Spark offer a fast and easy way to augment your data science capabilities with minimal investment in training, infrastructure and change management. With the IBM Analytics for Apache Spark service on Bluemix, you get the ability to integrate Spark with Cloudant data stores in minutes—all within the convenience of the Notebook toolset.

IBM Cloudant® + Spark

Case study:
Industrializing data science

Although many organizations are still taking the first steps on their data science journeys, it's wise to look further down the road and see how the most mature companies have successfully built in-house data science capabilities.

One IBM client has a large data science department, where each individual may be working on several projects—and multiple datasets—at any given time. Managing the lifecycles of all these different datasets was threatening to become an administrative nightmare, because it was difficult to tell whether a given dataset was actually in active use, or if it was part of an old project and could potentially be archived.

The company solved the problem by giving its users access to Cloudant, which they can use to create as many individual data repositories as they need. The change feeds of each repository are regularly sent to Spark through the Spark-Cloudant Connector, where they can be analyzed to reveal which datasets are currently in active use, and by which users.

The solution also helps make life significantly easier for data scientists who need to join data from multiple datasets together, or enrich a dataset with information from an external source. The Spark-Cloudant Connector can simply load the relevant data into separate Spark resilient distributed datasets (RDDs), perform the required joins or other transformations, and return the results for consumption. While Notebooks already offer a wide variety of built-in reporting and graphing capabilities, results can also be persisted for further downstream processing. A Cloudant database itself could be the target database for further analysis.

The bottom line: *Cloudant and Spark help make it easy for the company's data scientists to manage and analyze their data, avoiding a fragmented data landscape and facilitating the resuse and enrichment of important datasets.*

4 For the developer:

How can Cloudant and Spark help me innovate?

As organizations come to realize that innovation is the key to competitive advantage, the role of software developers is changing. Instead of being considered a backroom support function, the development team can potentially become a key driver of change.

In this new role, whenever an opportunity arises to build a solution that will help the business enter a new market or offer a new service, the development team needs to be able to take the initiative. Instead of asking for a large project team, a huge budget and an 18-month delivery cycle, it needs to be agile enough to create proofs-of-concept within days and potentially deliver a minimum viable product within weeks.

To make this possible, developers must have the right tools and technologies at their disposal. In many cases, traditional relational databases and on-premises application servers simply won’t allow for the kind of rapid prototyping and iterative development required. If your project starts out with procuring, installing and configuring a development server as a first step, it’s going to take weeks before you can even start coding in earnest—and by that time, your competitors may already be halfway to market.

Many of the most exciting opportunities for innovation now involve harnessing resources that are available in the cloud—whether that means capturing data from connected devices via the Internet of Things, channeling streams of social media data, or enriching internal data sources with external data feeds such as weather information.

For this reason, cloud-native technologies will typically be the best and fastest way to take advantage of these opportunities. Why build an on-premises relational database to download, transform and store non-relational data from the cloud, when you could simply utilize a NoSQL database that lives in the cloud and can handle that data natively, via developer-friendly APIs? And why purchase expensive, enterprise-class servers to process that data, when you could get the same result faster and more easily from a Spark service in the cloud? Taking advantage of cloud-based data and analytics services for these use cases dramatically helps reduce the complexity of software development projects—and therefore helps reduce the cost, risk, and delivery time as well.

For the developer:

How can Cloudant and Spark help me innovate?

So why use Cloudant and IBM Analytics for Apache Spark specifically? Aside from their easy availability within the Bluemix cloud application development environment, these two technologies offer a number of specific technical advantages to developers:

1. Flexible data storage
As a schema-less JSON document store, Cloudant minimizes the amount of data modeling and planning that developers need to conduct at the start of a project. Instead of designing and maintaining complex webs of tables and relationships, developers can store all their information in simple JSON data structures, enabling them to focus on developing new features rather than administering and optimizing the database.

Cloudant's JSON data model also makes it much easier to handle unstructured and semi-structured data types, and deal with sparse datasets and missing data. Without the rigid constraints of a relational database, it is easier to adapt to meet emerging requirements, or even pivot the entire project if necessary.

2. Easy data integration and transformation
Spark makes it easy to integrate Cloudant data stores with your existing landscape—you can use it as a super-fast data transformation engine to convert Cloudant JSON data to and from whatever formats your other systems need.
The wide range of connectors available for Spark (including the Spark-Cloudant Connector) mean that you can easily extract data from one system and load it into another, while Spark's rich set of transformation functions make it possible to turn one RDD with an inferred schema into another RDD with a defined schema, or vice versa.

As a result, Spark can be used to implement extract, transform load (ETL) applications and move data in and out of Cloudant very quickly. It can even be used to cleanse, consolidate and filter data in-place, by using a single Cloudant database as both the source and the target for the Spark-Cloudant Connector.

3. Offline-first application development
Cloudant Sync solves one of the classic problems of mobile application development: how to synchronize data efficiently between a local device and a central server. With Cloudant, seamless online and offline replication comes as standard and is available out-of-the box.
A well-tested use case is to use Cloudant as a staging area to land incoming data from mobile devices, and then use Spark to transform the data into whatever schema the back-end database platform requires—effectively slashing the cost and complexity of adding a mobile capability to an existing back-end system.

For the developer:

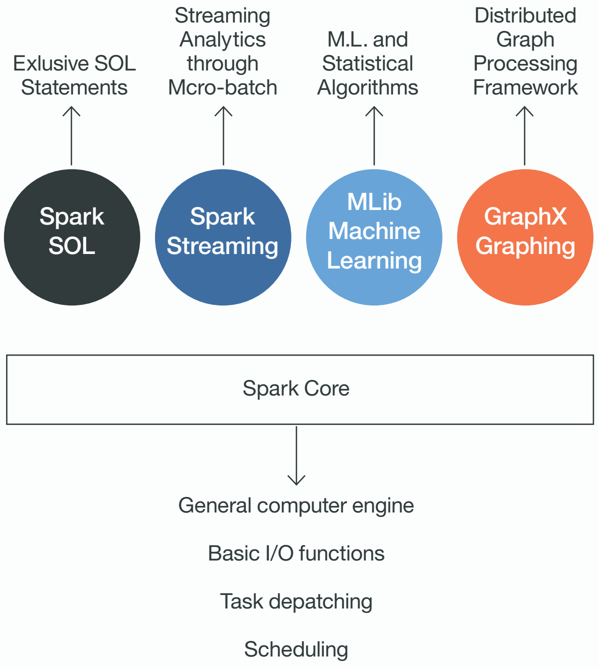
How will Cloudant and Spark help me innovate?

4. Geospatial capabilities
Cloudant offers native support for the GeoJSON standard, which makes it easy to develop sophisticated location-aware applications. The Spark-Cloudant Connector preserves the GeoJSON format, which means that Spark can be used to run complex geospatial analytics in near-real time—supporting map visualization, geo-clustering, route-finding, proximity analysis, and many other types of query.

5. Federated data processing
In many cases, today's cloud-enabled applications need to combine data from more than one source. The Spark-Cloudant Connector enables you to load individual documents, groups, or entire databases from multiple Cloudant sources into Spark RDDs and perform rapid joins even on the largest datasets. This can deliver a significant performance boost for complex queries with many joins.

6. Access to a broader analytics ecosystem
Cloudant is just one of the many data sources and other technologies that IBM is integrating with Spark. The entire ecosystem is already extensive, and is growing all the time—as you can see from Figure 1.

As a result, it's easy to create applications that mix and match a huge range of sources and services. You could combine JSON data from Cloudant, unstructured text from Hadoop, database tables from IBM dashDB™, and streaming data from Twitter; you could join these datasets and process them with machine learning libraries in Spark; and you could even draw in cognitive services such as natural language processing from IBM Watson™ APIs. The possibilities are almost endless—and with Bluemix as a shared development hub to help you coordinate and orchestrate these components, you can pull together a solution in days or weeks, that might have previously taken months or years to build from scratch.

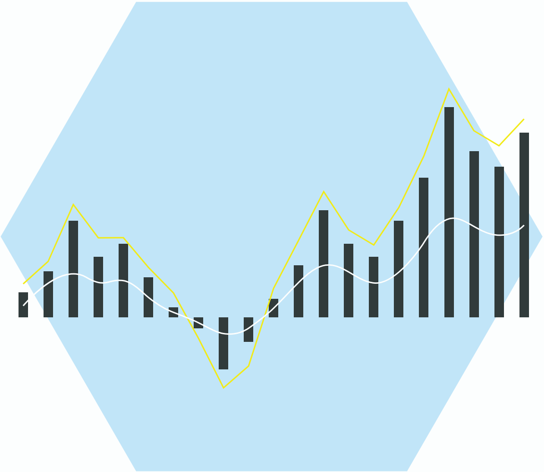


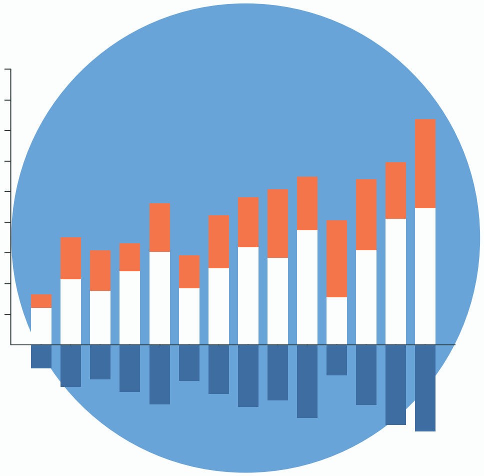
Case study:
Combining cloud data services

One IBM client, a network management company, realized it had an opportunity to learn more about how people move through public spaces (such as malls, hotels and airports) by analyzing the strength of the signals its wi-fi hotspots receive from their mobile phones. This helps the client it make smarter investments in new capacity by understanding which areas have the greatest demand for wi-fi services, and installing new hotspots in the right places. With thousands of hotspots, each logging several thousand connections with users' devices per day, this was truly a big data challenge.

The company now uses Cloudant as a master database to capture all of the connection records, grouped by hotspot. Spark is then used to analyze the details of each connection, and load the results into the company's dashDB cloud data warehouse, where it can be used for operational reporting and decision-making.

The bottom line: *Cloudant and Spark help developers orchestrate a complex, large-scale solution to a big data challenge, and give IT decision-makers a clearer view of how to plan the company's future investments.*





5 Take the next steps

If you are a data scientist or a developer, and you are already using Cloudant and/or Apache Spark, the Spark-Cloudant Connector can be found on GitHub or the Spark Packages site, and is available for free under the Apache 2.0 License. There are guided tutorials for using the connector with Python and Scala on IBM developerWorks.

As a data scientist, if you aren't using either technology yet, but would like to learn more, we recommend signing up for IBM Data Science Experience, where you can start using Spark and Cloudant with Jupyter Notebooks within minutes.

As a developer, the best way to start experimenting with Cloudant, Spark and a vast range of other cloud-data services is Bluemix. You can sign up for a free trial and see for yourself how easy it is to build sophisticated applications by linking together multiple different components via IBM's prebuilt integration frameworks. You might also want to check out this IBM webinar on how to build analytic apps with Cloudant.

If you are an IT decision-maker, you may want to take a look at some of IBM's case studies on Cloudant and Spark, to gain more insight into the potential use cases and benefits of these powerful technologies. For example, SolutionInc and the SETI Institute are both using Spark to address big data challenges around sorting signals from noise; while Quetzal and YukonBaby are harnessing Cloudant to solve application development challenges and quickly bring new products to market.

Finally, for a deeper dive into the world of big data analytics and data science with Spark and Cloudant, sign up at IBM Big Data University for free online courses to help you learn new skills and maximize your potential.

Ready to start analyzing your JSON Data?

Try out the Spark-Cloudant Connector ›

IBM Corporation
Commerce Group
Route 100
Somers, NY 10589

Produced in the United States of America
December 2016

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates. THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

[1] The spark.apache.org website mentions that Spark “[runs] programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk”, using the example of a logistic regression job that took 110 seconds on Hadoop and 0.9 seconds on Spark (webpage retrieved 6th October 2016).