Big data problems are often complex to analyze and solve. The sheer volume, velocity, and variety of the data make it difficult to extract information and business insight. A good first step is to classify the big data problem according to the format of the data that must be processed, the type of analysis to be applied, the processing techniques at work, and the data sources for the data that the target system is required to acquire, load, process, analyze and store.

Overview

Big data can be stored, acquired, processed, and analyzed in many ways. Every big data source has different characteristics, including the frequency, volume, velocity, type, and veracity of the data. When big data is processed and stored, additional dimensions come into play, such as governance, security, and policies. Choosing an architecture and building an appropriate big data solution is challenging because so many factors have to be considered.

This "Big data architecture and patterns" series presents a structured and pattern-based approach to simplify the task of defining an overall big data architecture. Because it is important to assess whether a business scenario is a big data problem, we include pointers to help determine which business problems are good candidates for big data solutions.

From classifying big data to choosing a big data solution

Try out IBM big data solutions

Download a trial version of an IBM big data solution and see how it works in your own environment. Choose from several products:

- **BigInsights Quick Start Edition**, the IBM Hadoop-based offering that extends the value of open source Hadoop with functions like Big SQL, text analytics, and BigSheets
If you've spent any time investigating big data solutions, you know it's no simple task. This series takes you through the major steps involved in finding the big data solution that meets your needs.

We begin by looking at types of data described by the term "big data." To simplify the complexity of big data types, we classify big data according to various parameters and provide a logical architecture for the layers and high-level components involved in any big data solution. Next, we propose a structure for classifying big data business problems by defining atomic and composite classification patterns. These patterns help determine the appropriate solution pattern to apply. We include sample business problems from various industries. And finally, for every component and pattern, we present the products that offer the relevant function.

Part 1 explains how to classify big data. Additional articles in this series cover the following topics:

- Defining a logical architecture of the layers and components of a big data solution
- Understanding atomic patterns for big data solutions
- Understanding composite (or mixed) patterns to use for big data solutions
- Choosing a solution pattern for a big data solution
- Determining the viability of a business problem for a big data solution
- Selecting the right products to implement a big data solution

### Classifying business problems according to big data type

Business problems can be categorized into types of big data problems. Down the road, we'll use this type to determine the appropriate classification pattern (atomic or composite) and the appropriate big data solution. But the first step is to map the business problem to its big data type. The following table lists common business problems and assigns a big data type to each.

#### Big data business problems by type

<table>
<thead>
<tr>
<th>Business problem</th>
<th>Big data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilities: Predict power consumption</td>
<td>Machine-generated data</td>
<td>Utility companies have rolled out smart meters to measure the consumption of water, gas, and electricity at regular intervals of one hour or less. These smart meters generate huge volumes of interval data that needs to be analyzed. Utilities also run big, expensive, and complicated systems to generate power. Each grid includes sophisticated sensors that monitor voltage, current, frequency, and other important operating characteristics. To gain operating efficiency, the company must monitor the data delivered by the sensor. A big data solution can analyze power generation (supply) and power consumption (demand) data using smart meters.</td>
</tr>
<tr>
<td>Category</td>
<td>Data Sources</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Telecommunications: Customer churn analytics</td>
<td>Web and social data, Transaction data</td>
<td>Telecommunications operators need to build detailed customer churn models that include social media and transaction data, such as CDRs, to keep up with the competition. The value of the churn models depends on the quality of customer attributes (customer master data such as date of birth, gender, location, and income) and the social behavior of customers. Telecommunications providers who implement a predictive analytics strategy can manage and predict churn by analyzing the calling patterns of subscribers.</td>
</tr>
<tr>
<td>Marketing: Sentiment analysis</td>
<td>Web and social data</td>
<td>Marketing departments use Twitter feeds to conduct sentiment analysis to determine what users are saying about the company and its products or services, especially after a new product or release is launched. Customer sentiment must be integrated with customer profile data to derive meaningful results. Customer feedback may vary according to customer demographics.</td>
</tr>
<tr>
<td>Customer service: Call monitoring</td>
<td>Human-generated</td>
<td>IT departments are turning to big data solutions to analyze application logs to gain insight that can improve system performance. Log files from various application vendors are in different formats; they must be standardized before IT departments can use them.</td>
</tr>
<tr>
<td>Retail: Personalized messaging based on facial recognition and social media</td>
<td>Web and social data, Biometrics</td>
<td>Retailers can use facial recognition technology in combination with a photo from social media to make personalized offers to customers based on buying behavior and location. This capability could have a tremendous impact on retailers’ loyalty programs, but it has serious privacy ramifications. Retailers would need to make the appropriate privacy disclosures before implementing these applications.</td>
</tr>
<tr>
<td>Retail and marketing: Mobile data and location-based targeting</td>
<td>Machine-generated data, Transaction data</td>
<td>Retailers can target customers with specific promotions and coupons based location data. Solutions are typically designed to detect a user’s location upon entry to a store or through GPS. Location data combined with customer preference data from social networks enable retailers to target online and in-store marketing campaigns based on buying history. Notifications are delivered through mobile applications, SMS, and email.</td>
</tr>
<tr>
<td>FSS, Healthcare: Fraud detection</td>
<td>Machine-generated data, Transaction data</td>
<td>Fraud management predicts the likelihood that a given transaction or customer account is experiencing fraud. Solutions analyze transactions in real time and generate recommendations for immediate action, which is critical to stopping third-party fraud, first-party fraud, and deliberate misuse of account privileges.</td>
</tr>
</tbody>
</table>
Solutions are typically designed to detect and prevent myriad fraud and risk types across multiple industries, including:

- Credit and debit payment card fraud
- Deposit account fraud
- Technical fraud
- Bad debt
- Healthcare fraud
- Medicaid and Medicare fraud
- Property and casualty insurance fraud
- Worker compensation fraud
- Insurance fraud
- Telecommunications fraud

Categorizing big data problems by type makes it simpler to see the characteristics of each kind of data. These characteristics can help us understand how the data is acquired, how it is processed into the appropriate format, and how frequently new data becomes available. Data from different sources has different characteristics; for example, social media data can have video, images, and unstructured text such as blog posts, coming in continuously.

We assess data according to these common characteristics, covered in detail in the next section:

- The format of the content
- The type of data (transaction data, historical data, or master data, for example)
- The frequency at which the data will be made available
- The intent: how the data needs to be processed (ad-hoc query on the data, for example)
- Whether the processing must take place in real time, near real time, or in batch mode.

**Using big data type to classify big data characteristics**

It's helpful to look at the characteristics of the big data along certain lines — for example, how the data is collected, analyzed, and processed. Once the data is classified, it can be matched with the appropriate big data pattern:

- Analysis type — Whether the data is analyzed in real time or batched for later analysis. Give careful consideration to choosing the analysis type, since it affects several other decisions about products, tools, hardware, data sources, and expected data frequency. A mix of both types may be required by the use case:
  - Fraud detection; analysis must be done in real time or near real time.
  - Trend analysis for strategic business decisions; analysis can be in batch mode.
- Processing methodology — The type of technique to be applied for processing data (e.g., predictive, analytical, ad-hoc query, and reporting). Business requirements determine the appropriate processing methodology. A combination of techniques can be used. The choice of processing methodology helps identify the appropriate tools and techniques to be used in your big data solution.
- Data frequency and size — How much data is expected and at what frequency does it arrive. Knowing frequency and size helps determine the storage mechanism, storage format, and the necessary preprocessing tools. Data frequency and size depend on data sources:
  - On demand, as with social media data
• Continuous feed, real-time (weather data, transactional data)
• Time series (time-based data)
• Data type — Type of data to be processed — transactional, historical, master data, and others. Knowing the data type helps segregate the data in storage.
• Content format — Format of incoming data — structured (RDMBS, for example), unstructured (audio, video, and images, for example), or semi-structured. Format determines how the incoming data needs to be processed and is key to choosing tools and techniques and defining a solution from a business perspective.
• Data source — Sources of data (where the data is generated) — web and social media, machine-generated, human-generated, etc. Identifying all the data sources helps determine the scope from a business perspective. The figure shows the most widely used data sources.
• Data consumers — A list of all of the possible consumers of the processed data:
  • Business processes
  • Business users
  • Enterprise applications
  • Individual people in various business roles
  • Part of the process flows
  • Other data repositories or enterprise applications
• Hardware — The type of hardware on which the big data solution will be implemented — commodity hardware or state of the art. Understanding the limitations of hardware helps inform the choice of big data solution.

depicts the various categories for classifying big data. Key categories for defining big data patterns have been identified and highlighted in striped blue. Big data patterns, defined in the next article, are derived from a combination of these categories.
Big data classification

Analysis Type
- Real Time
- Batch

Processing Methodology
- Predictive Analysis
- Hadoop
- Feature
- Time
- Statistical
- Transcription
- Speech
- Text
- 3D

Data Frequency
- On demand
- Continuous
- Real time
- Time series

Data Type
- Meta Data
- Master Data
- Historical
- Transactional

Content Format
- Structured
- Unstructured
- Semi-structured
- Images
- Text
- Videos
- Documents
- Audio

Data Sources
- Web and Social Media
- Machine generated
- Human generated
- Internal Data Sources
- Transactional Data
- Semantic Data
- Via Data Providers
- Via Data Originator

Data Consumers
- Human
- Business Process
- Other Enterprise Applications
- Other Data Repositories

Hardware
- Commodity Hardware
- State of Art Hardware

Conclusion and acknowledgements

Learn more. Develop more. Connect more.

The new developerWorks Premium membership program provides an all-access pass to powerful development tools and resources, including 500 top technical titles for application developers through Safari Books Online, deep discounts on premier developer events, video replays of recent O'Reilly conferences, and more. Sign up today.

In the rest of this series, we'll describe the logical architecture and the layers of a big data solution, from accessing to consuming big data. We will include an exhaustive list of data sources, and introduce you to atomic patterns that focus on each of the important aspects of a big data solution. We'll go over composite patterns and explain how atomic patterns can be combined to solve a particular big data use case. We will conclude the series with some solution patterns that map widely used use cases to products.

The authors would like to thank Rakesh R. Shinde for his guidance in defining the overall structure of this series, and for reviewing it and providing valuable comments.
Related topics

- Big Data University
- IBM Big Data and Analytics Hub
- Blog: Business Analytics 3.0: Analytics as a service
- Big data blog: James Kobielus
- InfoSphere Streams Quick Start Edition

© Copyright IBM Corporation 2013

Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)