Build a Scalable Platform for High-Performance IoT Applications

Abstract

Scalability will be key to handling the explosive growth in the Internet of Things (IoT). This means that IoT applications must have the ability to support an increasing number of connected devices, users, application features, and analytics capabilities, without any degradation in the quality of service. Scalable IoT applications are also essential to monitoring, securing, and managing an increasing number of devices through a proportionate increase in the resources.

There are several questions that must be answered in order to ensure the design of scalable IoT applications:

- What will happen if the frequency of data collection and the number of devices were to increase by a factor of ten or even one hundred?
- What kind of interactive queries are likely to be executed by end users? What will happen if the sizes of databases keep increasing?
- What kind of real-time requirements does the application have? Are there any deadline driven processing needs, and what if the deadlines are not met? Will the ability to meet the deadline suffer as the number of devices increases or if the frequency of data collection changes?

Armed with answers to these questions, businesses can identify ways of improving the scalability of their IoT applications.

About the Author

Prateep Misra

Prateep Misra is head of Technology and Engineering within the IoT Group at Tata Consultancy Services (TCS). He has over 23 years’ experience in the IT industry in areas such as software development, research, technology consulting, and software quality assurance. As a thought leader in TCS’ Innovation Labs, Prateep has been responsible for setting up Centers of Excellence (CoEs) in areas such as stream processing, open storage, RFID, and digital signal processing. His areas of expertise include embedded systems, RFID, IT infrastructure, real-time analytics, and cloud computing. He holds a BTech in Instrumentation Engineering and an MTech in Electrical Engineering, both from the Indian Institute of Technology, Kharagpur.

Five Key Strategies for Enhancing the Scalability of IoT Applications

An IoT application consists of functional components deployed in three distinct tiers: the edge, the platform, and the enterprise. The edge tier consists of sensors, edge devices, IoT gateways, sensor networks, and connectivity to the core via access networks. The platform tier comprises middleware, servers, storage, and core services needed for device management, data management, real-time processing, analytics engines, and so on. Typically these are hosted in a corporate data center or a cloud. The enterprise tier consists of applications and systems that make use of services provided by the platform tier.

Here we focus on the platform tier to look at various ways of improving scalability.

1. Use automated bootstrapping

When the number of devices increases, any manual process with respect to device bootstrapping, registration, software configuration, security configuration, and upgrade is no longer viable. Any process that requires a human operator to install, configure, deploy, and manage devices is also not feasible. The device software and management service in the platform tier must enable automated bootstrapping, registration, monitoring, and upgrade. Devices must come equipped with the necessary bootloaders, security keys, and other features. This will enable the device to initiate a sequence of steps when it powers up at a remote location for the first time. These sequential steps will allow it to securely identify and register itself with a server and request the server to load additional software as required.
2. Control the IoT data pipeline

IoT applications require a data processing pipeline consisting of front-end data collectors and a set of data curation (cleaning, enrichment, transformation) functions applied on streaming data. Data loggers that send the data to destinations such as databases, files, queues, and applications also play an important role. The data pipeline needs to be well balanced in its ability to sustain a continuous flow of streaming data. It must also be able to handle situations such as a temporary surge in the input rate, performance issues, and disruptions in downstream systems.

Capacity planning of such pipelines should be based on parameters such as the number of simultaneous connections or data streams, worst case input rates, and processing times, as well as the size of data packets. Understanding these parameters can help determine and provision the infrastructure needed at each stage of the pipeline. Furthermore, using the right tools to monitor resource usage, queue lengths, and response time, it is possible to automatically detect overloaded components and trigger the provisioning of additional resources. When a certain part of the pipeline is down or severely degraded, the data stream needs to be buffered until the pipeline becomes manageable. Buffers also help manage situations when there is a temporary input surge or mismatch between the input and output at any stage in the pipeline. Additionally, the pipeline may impose an input rate control to manage and streamline the pipeline.

3. Apply the three-axis scaling approach

IoT applications can be scaled along all three dimensions – the X, Y, and Z axes.

Scaling along the X axis involves using more resources to divide the incoming requests between multiple servers, so that any server can handle requests. This is also called horizontal scalability and is typically achieved by having a load balancer in front of the resource server. There are certain prerequisites that the application must satisfy in order to enable X-axis scaling. It is useful to employ servers that are ‘stateless’ - servers that do not internally maintain state information from one request to the next – since they are easier to scale and load-balance.

Scaling along the Y axis consists of dividing the workload based on actions. For example, ‘read resource’ requests may be sent to one set of servers and ‘update resource’ requests to another set.

Z-axis scaling is a way of dividing responsibilities based on the incoming request data or the response data. For instance, all devices belonging to a particular geographical region may be managed by one particular server. Y and Z axis scaling can be achieved by specialized routers or service gateways. ‘Consistent hashing’ is one technique by which such routing can be done.

4. Develop microservices architecture

In order to scale IoT applications, especially in the platform tier, it is useful to break down each application into multiple independent functional units, each of which performs one dedicated function. Each of these functional units should be independently deployable and executed. The functional units can send messages to each other. This style of architecture is called microservices architecture.

In an IoT application, major functional blocks such as device and data management, data stream processing and enrichment, and event processing should be broken down into one or more microservices. The overall application would combine these microservices through an orchestration logic. Moreover, these microservices, if designed properly, can also enable three-axis scaling. Automated monitoring agents and load balancers automatically spin up new instances of microservices as per the system load.

5. Adopt multiple data storage technologies

A ‘one size fits all’ approach may not work with respect to the choice of technology, especially for IoT data storage. Different parts of an IoT application such as the user interfaces, interactive querying, streaming analytics, batch oriented analytics, and machine learning algorithms must be built using the best-suited technology components. Data query and retrieval requirements, coupled with the analytics algorithms that run on the selected data, should determine the choice of data storage or the database technology. Each microservice should use a component that is ideally suited for its need.

For example, a set of data may persist in real time in the memory cache during stream processing. However, for model building applications, it might also be stored in a specialized array database. Similarly, for interactive querying with analytics, data might be stored in a distributed column oriented database. Such an approach ensures optimal performance and a highly responsive user interface. Basically, the user experience determines the entire architecture. While applying this approach to data storage, it is important to manage replication across different data stores.
Ensuring the long-term success of IoT applications

Performance, real-time requirements, security, and user experience, as well as the ability to manage, monitor, and maintain devices, are of immense importance for the success of any IoT application.

It’s obvious that the number of devices and applications will continue to grow. Gartner estimates that there will be 20.7 billion connected things by 2020.¹ As device functionality evolves and networks mature, it is imperative for businesses to enhance the scalability of their IoT applications to accommodate future growth and changing technology requirements. A well-engineered IoT platform can help businesses create a flexible architecture for developing elastic IoT applications that can support the ever-increasing number of devices, users, and data volumes.

About TCS’ Innovation Labs

Established in 1981, TCS’ Innovation Labs address real-world business problems, bringing scientific rigor to the study of computational concepts, and delivering solutions that make computation more reliable, efficient, and agile for our clients.

With a focus on three key aspects—Software, Systems, and Applications Sciences—our Innovation Labs build on new ideas to create solution frameworks that deliver high business impact across domains. Our Co-Innovation Network (COIN™) gives our clients access to potentially disruptive technologies through our partnerships with emerging technology companies, venture funds, academic research units, multi-lateral bodies, and Tata Group companies.

The global network of Innovation Labs is equipped with sophisticated infrastructure to support leading-edge research. Our researchers file over a hundred patents each year, and our research-based papers are a recognized voice on global thought leadership platforms. Our breakthroughs have won several industry awards, including the MIT Technology Grand Challenge Award and the Stockholm Challenge Award.

Contact
Visit TCS’ Innovation Labs page for more information
Email: tcs.iot@tcs.com
Blog: Research and Innovation

About Tata Consultancy Services Ltd (TCS)
Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match.
TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India’s largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com