Data Collection & Analysis

Usage and Capacity Planning White Paper
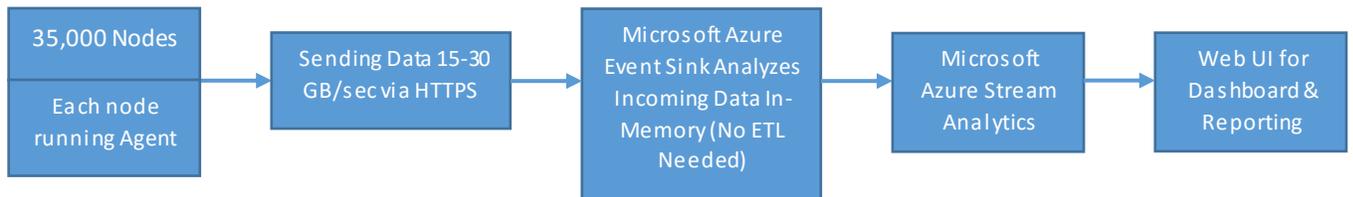Terabytes of In-Memory Analysis – No ETL – Ever!

Overview

Our customer wanted a deeper understanding of the usage of its IoT resources with over 35,000 installations. Getting an insight into the usage will allow our customer to better plan for capacity and ensure resources were available to serve our customer's needs. Our client wanted to collect the details of each node, most heavily used items within each node, health of the node and behavior of the node so that collected data can predict health, quality and overall capacity of the system.

The approach to gathering the data is to record events as the node is used and then periodically send those events to some back-end solution for storage and analysis. This white paper will provide details on how BPK Tech Engineering Division helped develop a solution for the event aggregation, analytics, and storage.

Solution

We built a solution by reviewing the infrastructure, location of nodes, building nodes APIs, communicating to Microsoft Azure Event Sink, analyzing data on the Stream Analytics and the building a web application to present results to decisions makers. It is depicted below:

35,000 Nodes Each node running Agent	→	Sending Data 15-30 GB/sec via HTTPS	→	Microsoft Azure Event Sink Analyzes Incoming Data In-Memory (No ETL Needed)	→	Microsoft Azure Stream Analytics	→	Web UI for Dashboard & Reporting

There is no need to do any ETL, ever!

Stream Analytics will do aggregations and store the raw input as blobs in blob storage and the aggregated data in Azure Table Storage.

To view the results of the analytics, a very basic web application was setup to allow business users to log in and view the results. For use cases where the results are grouped by network strength or location or customer or router, the web application will allow for searching/filtering.

Security

Security was the cornerstone for this application – we employed multiple levels of security, one for listeners and one for senders. The shared access key for senders and listener was done via Token Service (SAS Token). When publishers send events, they will be expected to use an SAS token obtained from the token service. The actual shared key will never be shared with publishers.

Traffic Volumes and Scaling

With over 35,000 nodes reporting data, all of which may have sub-nodes, and the fairly fine-grained events, it is reasonable to expect a significant volume of events flowing into the system. We expected 1.2 – 7.5 million events per hour (2,000 – 12,500 event per second) at peak times – however, we built the infrastructure in Azure to handle 12-75 million events per hour. The raw data ingress was closed at 20 GB per second.

Want us to solve a problem?

We can help you solve this or any other software development or analytics problem. We have done ton of such implementations – let us help. You can reach us at bettersystem@bpktech.com or (612) 293 7585.