Predictive Cloud Computing for professional golf and tennis, Part 1: Introduction

Aaron K. Baughman
Cameron N. McAvoy
Brian O’Connell
Nicholas A. McCrory

January 18, 2016

Major golf and Grand Slam tennis tournaments provide real-time and historical sporting information to immerse their fans in the action. Each tournament provides content that includes streaming video, game statistics, scores, images, schedule of play, and text. This tutorial describes how Predictive Cloud Computing (PCC) forecasts and predicts tournament popularity to automatically allocate shared computing resources as needed. PCC thus saves sporting customers many hours per day while displaying key statistics about the impact of sporting big data on the infrastructure.

To demonstrate the system of big data, analytics, and cloud, this tutorial presents Predictive Cloud Computing (PCC), which has been deployed to professional golf and tennis tournaments since 2012. This is the first in a series of tutorials that detail the PCC system with the goal of outlining best practices for developers. Part 1 focuses on web application coding and design on IBM WebSphere Liberty Profile servlet and software design of a RESTful-based forecasting engine.

“This tutorial describes how Predictive Cloud Computing (PCC) forecasts and predicts tournament popularity to automatically allocate shared computing resources as needed.”

Problem statement

The rapid growth and accessibility of digital content is driving the growth of the Internet and is producing an increased demand on enterprise-cloud computing resources. The patterns of large amounts of web traffic are dynamic and uncertain, yet always-on or continuously available cloud services must always meet demand. Optimal operational efficiencies allocate more cloud resources during an influx of traffic and deallocate resources when idle. The PCC project utilizes
a combination of analytics, big data, and cloud to effectively manage computing resources during professional golf and tennis tournaments.

In 2014, PCC supported six of the eight major events shown in Figure 1, which include 80 event competition days, in four countries, six websites, and six mobile sites. The cloud demand translated into 1.5+ billion page views, 191+ million visits, 60+ million unique visitors/devices, and 7.5+ million live scoring updates. The 2015 golf and tennis tournament season that IBM sponsored started in January with the Australian Open and ended in October with the China Open. The blistering schedule pace required agile development and continuous deployment.

Figure 1. The Events Infrastructure (EI) supports a blistering schedule of eight major tournaments and entertainment events

Overall architecture and design

Within PCC, hindsight, which is provided by the inspection of descriptive analytics, generates clear insights, whereas foresight, derived from forecasting and predictive modeling, enables proactive discernment. High-volume big data components such as sporting information, cloud data, and social streams are refined by analytics into knowledge. Each set of data is either at rest, in motion, or pushed through analytic pipelines that require specific big data architectures. Cloud computing provides computational resources on demand at the time of need to support big data and analytics.

Real-time web server access logs are streamed into PCC through a Python collection script. The access logs are parsed and put into JavaScript Object Notation (JSON). Each JSON message is sent to an exchange that forwards to an appropriate queue that matches a routing key. IBM® InfoSphere® Streams then consumes the log messages from the queue and aggregates each into minute time bands. IBM InfoSphere Streams also subscribes to Twitter's PowerTrack. Tweets are processed for social sentiment about particular players. Both the streaming logs and tweets are sent to a RESTful web service for analysis and storage into IBM DB2®. Gameplay information is consumed by a web application and saved into DB2 for future processing.
Figure 2 depicts high-volume data at rest that is stored within IBM InfoSphere BigInsights® Hadoop Distributed File System (HDFS). Web access logs are stored within HDFS and correlated with web crawler data that seeks mentions of players on web pages. As a result, player popularity is determined based on how often a player is mentioned on a site and the magnitude of accesses for the particular web page. The BigInsights job is managed by Oozie where the output of the job is stored within a relational database, DB2.

Figure 2 shows several types of analytics that consume the Big Data. Under "Analytics," the group of pre-processors accept real-time traffic information that is binned by the minute (step 1). The historical traffic data from previous days is appended and time shifted with the real-time traffic data. The pre-processors create seasonality curves and impute any missing values. The time-series ensemble applies five complementary forecasters of differing techniques to predict future server demand (step 2).

Next, the post-processors filter any numerical errors or duplicate values, time shift the forecast, remove anomalous forecasts, and smooth the cohort half-life weighted cyclical forecast curve.

At step 3, a distributed chained discrete event simulator runs to simulate golf or tennis tournaments into the future. A distributed feature extraction system runs on UIMA-AS and applies algorithms on the simulated game state, tweets, log traffic, and published sporting data (step 4). Step 5 depicts the process of chaining forward the simulated game state to run additional simulations for feature extraction. The resulting feature vector is used either to train or apply a multiple linear regression model (step 6). Large increases in traffic or spikes are detected by the predictive model at step 7 to produce an event forecast. A residual post-processor adjusts the forecast by the average mathematical errors from Powell Optimization, Loess Interpolation, and Loess Extrapolation (step 8). The cyclical forecast and event forecast are combined by a sliding parabolic weight adjuster at step 9. The resulting composite forecast adjusts the number of web servers within the cloud (step 10).
Figure 2. The PCC system uses analytics to interpret Big Data to automatically adjust the Cloud

Figure 3 depicts an overall component diagram across several networking zones. The Red Zone is the entry point for Internet users to access PCC through a web acceleration tier. Global Server Load Balancers (GSLB) utilize a global Domain Name Service (DNS) to balance traffic across a three-sited continuous availability cloud architecture. The GSLBs detect and respond to failures to provide continuity, resiliency, and accessibility of PCC.

The Yellow Zone provides a web tier that balances traffic within a specific site. IBM HTTP Servers (IHS) serve web access containers that in turn forward traffic to the application tier. In addition, Python scripts capture the web access logs and act as a producer for data into HDFS and RabbitMQ.

In the Green Zone, application containers provide services for consumption. IBM WebSphere Liberty Profile (WLP) serves a complex Web application ARchive (WAR) application written in Java™ 1.7 that provides forecasting and cloud provisioning insights. A database server runs several instances of DB2 10.5.5 that are accessed by UIMA-AS Java Virtual Machines (JVM), WLPs, and BigInsights (BI). IBM InfoSphere Streams consumes data from RabbitMQ on the green zone and Twitter GNIP through network flows.

Each of the red, yellow, and green zones is firewalled to prevent unauthorized entry. The event's infrastructure runs on any number of private clouds with Figure 3 depicting the minimum of three for continuous availability that are managed by SoftLayer or internal event infrastructure.
High and continuous availability

Cloud computing refers to on-demand and always-on computing resources that can be obtained through the Internet. Generally, two types of cloud-service-level agreements are defined: high availability (HA) and continuous availability (CA). High availability generally delivers cloud services with 99.99% or similar availability during scheduled periods. Continuous availability delivers no unplanned or planned outages, high availability, and continuous operations at the 99.999% service. The service can transparently withstand component failures and disasters while maintaining consistency.

The Event's Infrastructure provides continuous availability to IBM's eight sponsored sporting and entertainment events. To maintain CA, PCC cannot under-provision resources, which would increase the likelihood of a service outage. The proactive cloud resource management has to maintain accurate forecasts. Over all event days, the PCC has shown an impressive average Mean Absolute Percentage (MAP) error of ~10%. The CA requirements are mitigated with the high MAP as well as safeguards to ensure resources are never provisioned below a predetermined threshold.

Media references

During each of the IBM-sponsored events, we prominently feature PCC within a technology showcase that PCC initiates deeper conversations around IBM's Big Data and analytics capabilities. Select media outlets are invited to write articles about PCC. Computerworld, All
Things D, Forbes, Information Week, Power ITPro, ZDNet, and others have all published articles about our work. **Table 1** summarizes a few media accolades that are publicly available on the Internet.

### Table 1. PCC has received numerous accolades from the media

<table>
<thead>
<tr>
<th>Date Published</th>
<th>Publisher</th>
<th>Name of article and link</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/05/2013</td>
<td>Computerworld</td>
<td>10 intriguing real-world uses for big data</td>
</tr>
<tr>
<td>04/11/2013</td>
<td>All Things D</td>
<td>How IBM Brings the Masters to Golf Fans</td>
</tr>
<tr>
<td>04/12/2013</td>
<td>Power ITPro</td>
<td>How IBM Works with Masters to Deliver an Immersive Digital Experience</td>
</tr>
<tr>
<td>04/11/2014</td>
<td>Forbes</td>
<td>IBM At The Masters: A Sponsorship Unlike Any Other</td>
</tr>
<tr>
<td>08/30/2014</td>
<td>Information Week</td>
<td>US Open Tennis: 7 Technologies Power Game, Set, Match</td>
</tr>
<tr>
<td>04/09/2015</td>
<td>ZDNet</td>
<td>How IBM's predictive cloud makes the Masters' website virtually uncrashable</td>
</tr>
</tbody>
</table>

### IEEE and INFORMS achievements

The Institute for Electrical and Electronic Engineers (IEEE) publishes high-quality peer-reviewed articles with several leading scientific journals and magazines. The IEEE Computational Intelligence Society sponsors a magazine called *Computational Intelligence Magazine* (CIM). Despite an acceptance rate of less than 10%, CIM accepted our tutorial titled "Predictive Cloud Computing with Big Data: Professional Golf and Tennis Forecasting," published in 2015. The tutorial presents an empirical, metrics-driven approach for evaluating the design and efficacy of our algorithms.

**Watch Predictive Cloud Computing with Big Data**

Simultaneously, the Institute for Operations Research and Management Science (INFORMS) accepted our PCC entry into the prestigious Franz Edelman competition. Since 1952, the Franz Edelman competition has selected the most influential advanced analytic and operations research projects from around the world to compete for the Franz Edelman prize. In 2015, we competed with Ingram Micro, LMI/Defense Logistics Agency, Saudi Arabia Ministry of Municipal and Rural Affairs, Syngenta, and the US Army/Sandia National Laboratories. Our PCC work was a runner-up for the award and earned IBM a Franz Edelman trophy and placed it into the Edelman academy. Each member of the PCC team became Franz Edelman Laureates for life and earned an Edelman medal and certificate. On April 13, we defended our PCC work; a video recording of that defense is available. In early 2016, INFORMS's flagship journal will publish an article about PCC titled "IBM Predicts Cloud Computing Demand for Sports Tournaments" that details the business problem and impact.

### Technology impact

The technology impact of the PCC continues to advance, as does the predictive analytics and enterprise cloud market. The PCC drives more operational efficiencies, analytics offerings, and an
ability to open doors with potential customers. As depicted in a video, the PCC is transportable and valuable to many sectors within the global economy.

To view this video, Predictive Cloud Analytics Drives an Exceptional Fan Experience, please access the online version of the article. If this article is in the developerWorks archives, the video is no longer accessible.

In 2013 alone, we presented to over 17,000 people through four IBM conferences, including Pulse, Edge, Enterprise, and Information on Demand. In 2014, we were the keynote for a conference that included 50 North Carolina C-Level female executives, we were featured at IBM’s Technical Leadership Exchange in New York, and we presented at the IBM Global Finance Summit in Charleston, South Carolina. Over the past two years, we have participated in over 25 PCC-related commercial engagements.

Throughout the PCC project, we not only have published scientific and business articles, but substantially contributed to IBM’s intellectual property. Sixteen utility patents have been filed within the areas of advanced analytics, cloud computing, machine learning, forecasting, networks, simulation, social networking, and hyper parameters. The patents forge a path for IBM to continue innovating within the PCC space.

From an operational perspective, the PCC has saved 51 percent of our compute hours or 134 hours per day during each event. As a result, over the span of 80 competition days, the PCC saved 446.7 compute days. The compute cycles could then be allocated to other customers that required higher load.

IBM Cloud

Several components of PCC can run on IBM Cloud®, which is a cloud Platform as a Service based on Cloud Foundry that runs on the SoftLayer infrastructure. See the IBM Cloud catalog for a list of Cloud services. Figure 4 depicts Cloud services that can be used to support PCC.

Figure 4. IBM Cloud provides several cloud services that support web applications, programming environments, messaging, data storage, and analytics across the PCC

IBM Cloud

Several components of PCC can run on IBM Cloud®, which is a cloud Platform as a Service based on Cloud Foundry that runs on the SoftLayer infrastructure. See the IBM Cloud catalog for a list of Cloud services. Figure 4 depicts Cloud services that can be used to support PCC.

Figure 4. IBM Cloud provides several cloud services that support web applications, programming environments, messaging, data storage, and analytics across the PCC

The Liberty for Java and IBM Liberty services in IBM Cloud provide containers for IBM WebSphere® Liberty Profile that service the PCC BigEngine application. The Python Community
enables a Python environment to run PCC access log collection. The Monitoring and Analytics services augment the use of Graphite, which will be discussed in a future tutorial. RabbitMQ is supported with the CloudAMQP service. Several database services such DB2 on Cloud and Time Series Database can supply containers for the PCC DB2 database and the Graphite time series data store. The Hadoop-based IBM InfoSphere BigInsights runs on the Cloud service called BigInsights for Apache Hadoop. Additional insights from Twitter tweets are provided by the Cloud service Insights for Twitter, while the multivariate linear predictive model in the PCC can be deployed by the Predictive Modeling Cloud service.

Conclusion

In this tutorial, we showed how Predictive Cloud Computing is used during major sporting events to balance dynamic work loads over hybrid cloud-based resources and thus provide real-time information to sports fans around the world. We described the architecture used to apply descriptive analytics to high-volume Big Data components, both at rest and in motion, to provide insight into future infrastructure demands. We also explained how PCC provides continuous availability by enhancing IBM's real-time resource allocation across a global hybrid cloud infrastructure.

In part 2 of this series, we will describe IBM's use of WebSphere Liberty Profile and our BigEngine application to provide sporting tournament simulation based on social insights, predictive modeling, and time series forecasting. We will also provide detailed examples of IBM's use of Git, Urban Code Deploy, and Java tools such as Maven and Jenkins, to forecast near-future computing requirements for web applications receiving large numbers of simultaneous requests.
Related topics

- Explore IBM Cloud
- IBM Predicts Cloud Computing Demand for Sports Tournaments

© Copyright IBM Corporation 2016
Trademark
(www.ibm.com/developerworks/ibm/trademarks/)