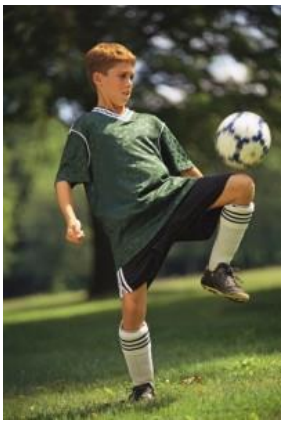
Can SAS Data Management get you to soccer on time?

By Matthew Magne

A soccer fairy tale



Imagine it's Soccer Saturday. You've got 10 kids and 10 loads of laundry – along with buried soccer jerseys – that you need to clean before the games begin. Oh, and you have two hours to do this. Fear not! You are a member of an advanced HOA (Home Owner Association) that provides an agile and flexible washing machine architecture (we'll call it AWA for Agile Washing Architecture) to improve the productivity of all your neighbors.

Whenever anyone in the neighborhood needs to wash a lot of laundry, they engage the AWA – and a large, driverless, Uber-like minivan pulls up alongside the house with 10 advanced upright washing machines and dryers in it (with steam control!). You tell your 10 kids to load up one load of laundry each into the 10 machines in the AWA. Two hours later, your family is on the way to 10 soccer games in an extended minivan, with fresh new soccer clothes. **Mission accomplished.** (Luckily, today they are all playing in the same park.)

When it comes to managing data, this isn't just a fairy tale about advanced HOAs and getting to soccer on time (though both of these scenarios are indeed fairy tales). It's an example of what happens when we can distribute processing into discrete chunks (loads of laundry) across nodes on an Apache Hadoop cluster (washing machine) that runs some sort of data cleansing, blending or transformation functions (cleaning the laundry). Instead of taking 10 hours to wash 10 loads, it takes 1 hour to wash (and 1 to dry).

Now, imagine your children placed their dirty laundry directly into the AWA washing machines at the end of each day. At that point, you could hit a button on the AWA to return only the specific subset of soccer clothes you needed – automatically – from the 10 loads of laundry.

That's an example of managing data where it lives (the driverless van comes to you versus you having to drive 5 miles to take your dirty clothes to a laundromat). And it illustrates the notion of bringing the processing to the data – which speeds processing (i.e., cleaning) and extracting only the data you need (i.e., soccer laundry items). In turn, you get faster performance and improved security.

SAS can help avoid common data quality issues

While our kids may not get to soccer on time, we've been busy over here at SAS making laundry loads of improvements to the SAS Data Management portfolio. SAS has improved the cleansing, preparation, streaming (sorry, not steaming) and virtualization capabilities of its data management suite. That includes SAS Data Loader for Hadoop, SAS Event Stream Processing and SAS Federation Server – which drive better analytics and manage your data right where it lives. So your organization will be able to deliver secure, streaming and streamlined data with an agile infrastructure that promotes improved efficiency, performance and accuracy.

Organizations can manage data where it lives and apply consistent data management rules across different environments while minimizing data movement and improving performance and governance. For example, the same data quality rules can be applied to data in motion inside the event stream using SAS Event Stream Processing, or inside Hadoop using SAS Data Loader for Hadoop. Or, the rules can be dynamically generated inside a virtual view of the data using SAS Federation Server.

SAS is one of the few innovators with an integrated suite of solutions that share metadata across both the data management and analytics domains. Competitors have a hodgepodge of acquisitions or partnerships with metadata stored in multiple places, either of which slows down integration and time to value.

Some benefits of SAS Data Management

- **Improved accuracy.** SAS can execute real-time, on-demand functions to fix common data quality issues. For example, it can parse your data, guess gender based on names, or de-duplicate data inside a virtualized view, in-memory using Apache Spark, or in-motion inside the data stream. And machine learning models can be applied to data in motion.
- **Higher performance and productivity.** Native Hadoop drivers, Impala queries (a faster way to run SQL on Hadoop), and in-database merging and execution speed performance. Exposed REST APIs ease external job scheduling. Models that analyze streaming data can be changed while data is in motion without taking the solution off-line.
- **Better security and streamlined integration.** Shared metadata spans both the analytics and data management domains for simpler integration. Dynamic data masking provides role-based ability to obfuscate data. Integration with 200+ data endpoints is simplified using Apache Camel adapters. And integration with the most prevalent Hadoop distributions is supported.

So, the next time you're on the way to soccer with last week's smelly jerseys on, think about this: Wouldn't it be great to have the power of AWA (and the power of SAS Data Management) in your driveway?