HCatalog, which is licensed under the Apache License, is a table management layer for the Apache Hadoop framework. It provides a table abstraction that enables you to think about data in the Apache Hadoop framework as if it were a set of tables. By eliminating the need to worry about the differences in file systems, where data is stored, its format, and which tools are required to translate to that format, HCatalog makes data in the Apache Hadoop framework available to users inside and outside the system.

HCatalog addresses one of the emerging issues with the Apache Hadoop framework today: With so many different tools to use, each with clear strengths for different use cases, how can you interoperate on the data?

These specialized tools are one of the great strengths of the Apache Hadoop framework. There’s no longer only one door to the data—through schema based on Structured Query Language (SQL). With the Apache Hadoop framework, you can pick the best tool for the job depending on your data format. However, because users of Apache Hive*, Pig, and MapReduce each think about their data in a different way, it’s difficult for them to share data together. As an integration tool, HCatalog enables interoperability through the table abstraction. It presents a common table layer to each tool and gives each user the same data model.

Consequently, users of these tools can all understand and interact with the data in the Apache Hadoop framework. For example, if you’re a Pig user, as soon as your Pig Latin script finishes and writes out its data, a Hive user will instantly see that as available as if it were Hive data. You no longer have to take that data and load it into Hive—it’s automatically shared across users.

Integration with Enterprise Data Management Tools

HCatalog also integrates with enterprise data management tools external to the Apache Hadoop framework through a Representational State Transfer (REST) interface, a language-independent transport that enables other systems to interact with metadata. Through REST, these systems can interrogate HCatalog for information about table schema, partitions, and other operations. External systems can also create or modify those tables, assuming they have the proper permission or create a new table so that your application can load data into it.
Apache Incubator* Projects

HCatalog is an Apache Incubator project, which means it's an effort in the early stages of development at the Apache Software Foundation (ASF). Sponsored by the Apache Incubator PMC, incubation is a process where new projects grow, learn the Apache way, and eventually graduate to become what are called top-level projects, for example Apache Hadoop and Apache Pig.

Being in the incubator does not mean the code is immature. Rather, the incubator is about developing community maturity and support for the code as part of overall Apache efforts. A project is in the incubator until the Apache Incubator PMC determines that the community is mature enough to be its own project. At that point, it gains full ASF endorsement.

Next Steps for HCatalog

HCatalog has a lot of room for growth. It's a new incubator project; HCatalog released its first stable release as version 0.4 in May 2012. We've already seen early interest from companies that want to roll it out into their environments, such as Yahoo! and Twitter. Additional growth areas include:

- **Strengthening the interface with external tools.** The community has made great progress with this via the REST interface, but there's more to be done. Defining data interchange formats between the Apache Hadoop framework and external systems capable of managing parallel streams is one area of development. In that case, you would be able to connect your Apache Hadoop system to another large multiprocessing database and pull data to help answer queries. Today you can't move that kind of data volume through a single open database connectivity (ODBC) connection.

- **Developing interfaces to emerging Apache Hadoop maintenance tools.** The Apache Hadoop framework doesn't yet have the kinds of maintenance tools you would expect of a big data processing system. Standard tools for cleaning a Hadoop* cluster, replicating data between clusters, and archiving data within a cluster all need to be developed. HCatalog will need the right interfaces to enable these tools.

- **Expand a greater number of use cases.** One of the strengths of the Apache Hadoop framework is its ability to manage semistructured or unstructured data. Because HCatalog presents a table interface, we're currently focused on data that can fit that structure. We need to work on handling use cases that include data that isn't as nicely structured, or where the structure may be unknown at load time and only discoverable at read time.

Statistical Support in HCatalog

HCatalog currently supports table-level statistics. Apache Hadoop* workloads are best suited for on-the-fly statistics gathering during processing rather than collecting them in the background to be stored in a relational database. More work needs to be done to support query optimizations for finer-grained statistics—for example, output of a histogram describing the distribution of values in a column.

This paper is for informational purposes only. THIS DOCUMENT IS PROVIDED “AS IS” WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION, OR SAMPLE. Intel disclaims all liability, including liability for infringement of any property rights, relating to use of this information. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

Copyright © 2012 Intel Corporation. All rights reserved. Intel, the Intel logo, Intel Sponsors of Tomorrow., and the Intel Sponsors of Tomorrow. logo are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.

This paper is derived from an interview with Alan Gates on August 28, 2012. For the full interview, listen to the podcast.

For more information about the Apache HCatalog project, visit [http://incubator.apache.org/hcatalog](http://incubator.apache.org/hcatalog).