I dedicate my work to my mother, Devjani Sarkar.

All that I am, or hope to be, I owe to you my Angel Mother. You have been my inspiration throughout my life. I learned commitment, responsibility, integrity and all other values of life from you. You taught me everything, to be strong and focused, to fight honestly against every hardship in life. I know that I could not be the best son, but trust me, each day when I wake up, I think of you and try to spend the rest of my day to do anything and everything just to see you more happy and proud to be my mother. Honestly, I never even dreamed of publishing a book some day. Your love and encouragement have been the fuel that enabled me to do the impossible. You’ve been the bones of my spine, keeping me straight and true. You’re my blood, making sure it runs rich and strong. You’re the beating of my heart. I cannot imagine a life without you, Love you so much MA!
# Contents

About the Author ................................................................. xiii  
About the Technical Reviewers .......................................... xv  
Acknowledgments ............................................................. xvii  
Introduction ........................................................................ xix  

- **Chapter 1: Introducing HDInsight** ........................................ 1  
  What Is Big Data, and Why Now? ............................................. 1  
  How Is Big Data Different? ..................................................... 3  
  Is Big Data the Right Solution for You? ................................. 3  
  The Apache Hadoop Ecosystem .............................................. 5  
  Microsoft HDInsight: Hadoop on Windows ......................... 7  
  Combining HDInsight with Your Business Processes ............. 9  
  Summary ............................................................................. 12  

- **Chapter 2: Understanding Windows Azure HDInsight Service** ........................................................................... 13  
  Microsoft’s Cloud-Computing Platform ................................. 13  
  Windows Azure HDInsight Service ....................................... 14  
    HDInsight Versions............................................................. 15  
    Storage Location Options ................................................. 16  
    Windows Azure Flat Network Storage .............................. 20  
  Summary ............................................................................. 22  

- **Chapter 3: Provisioning Your HDInsight Service Cluster** ...................................................................................... 23  
  Creating the Storage Account .............................................. 23  
  Creating a SQL Azure Database ......................................... 26
Chapter 6: Exploring the HDInsight Name Node ................................................................. 89
Accessing the HDInsight Name Node .................................................................................. 89
Hadoop Command Line ............................................................................................................ 92
    The Hive Console .................................................................................................................. 96
    The Sqoop Console .............................................................................................................. 97
    The Pig Console .................................................................................................................. 101
Hadoop Web Interfaces ......................................................................................................... 104
    Hadoop MapReduce Status ................................................................................................. 104
    The Name Node Status Portal ............................................................................................ 106
    The TaskTracker Portal ..................................................................................................... 107
HDInsight Windows Services ............................................................................................... 108
Installation Directory ........................................................................................................... 110
Summary ............................................................................................................................... 111

Chapter 7: Using Windows Azure HDInsight Emulator ..................................................... 113
Installing the Emulator .......................................................................................................... 114
Verifying the Installation ....................................................................................................... 116
Using the Emulator ............................................................................................................... 124
Future Directions .................................................................................................................. 125
Summary ............................................................................................................................... 125

Chapter 8: Accessing HDInsight over Hive and ODBC ..................................................... 127
Hive: The Hadoop Data Warehouse ..................................................................................... 127
Working with Hive .................................................................................................................. 129
    Creating Hive Tables ........................................................................................................... 129
    Loading Data ....................................................................................................................... 134
    Querying Tables with HiveQL ............................................................................................. 135
Hive Storage .......................................................................................................................... 137
The Hive ODBC Driver .......................................................................................................... 137
    Installing the Driver ............................................................................................................ 137
Sqoop Jobs

Windows Azure Storage Blob
  WASB Authentication
  Azure Throttling

Connectivity Failures

Summary

Index
Debarchan Sarkar (@debarchans) is a Senior Support Engineer on the Microsoft HDInsight team and a technical author of books on SQL Server BI and Big Data. His total tenure at Microsoft is 6 years, and he was with SQL Server BI team before diving deep into Big Data and the Hadoop world. He is an SME in SQL Server Integration Services and is passionate about the present-day Microsoft self-service BI tools and data analysis, especially social-media brand sentiment analysis. Debarchan hails from the “city of joy,” Calcutta, India and is presently located in Bangalore, India for his job in Microsoft’s Global Technical Support Center. Apart from his passion for technology, he is interested in visiting new places, listening to music—the greatest creation ever on Earth—meeting new people, and learning new things because he is a firm believer that “Known is a drop; the unknown is an ocean.” On a lighter note, he thinks it’s pretty funny when people talk about themselves in the third person.
About the Technical Reviewers

Rodney Landrum went to school to be a poet and a writer. And then he graduated, so that dream was crushed. He followed another path, which was to become a professional in the fun-filled world of Information Technology. He has worked as a systems engineer, UNIX and network admin, data analyst, client services director, and finally as a database administrator. The old hankering to put words on paper, while paper still existed, got the best of him, and in 2000 he began writing technical articles, some creative and humorous, some quite the opposite. In 2010, he wrote The SQL Server Tacklebox, a title his editor disdained, but a book closest to the true creative potential he sought; he wanted to do a full book without a single screen shot. He promises his next book will be fiction or a collection of poetry, but that has yet to transpire.

Scott Klein is a Microsoft Data Platform Technical Evangelist who lives and breathes data. His passion for data technologies brought him to Microsoft in 2011 which has allowed him to travel all over the globe evangelizing SQL Server and Microsoft’s cloud data services. Prior to Microsoft Scott was one of the first 4 SQL Azure MVPs, and even though those don’t exist anymore, he still claims it. Scott has authored several books that talk about SQL Server and Windows Azure SQL Database and continues to look for ways to help people and companies grok the benefits of cloud computing. He also thinks “grok” is an awesome word. In his spare time (what little he has), Scott enjoys spending time with his family, trying to learn German, and has decided to learn how to brew root beer (without using the extract). He recently learned that data scientists are “sexy” so he may have to add that skill to his toolbelt.
Acknowledgments

This book benefited from a large and wide variety of people, ideas, input, and efforts. I’d like to acknowledge several of them and apologize in advance to those I may have forgotten—I hope you guys will understand.

My heartfelt and biggest THANKS perhaps, is to Andy Leonard (@AndyLeonard) for his help on this book project. Without Andy, this book wouldn’t have been a reality. Thanks Andy, for trusting me and making it possible for me to realize my dream. I truly appreciate the great work you and Linchpin People are doing for the SQL Server and BI community, helping SQL Server to be a better product each day.

Thanks to the folks at Apress, Ana and Jonathan for their patience; Roger for his excellent, accurate, and insightful copy editing; and Rodney and Scott for their supportive comments and suggestions during the author reviews.

I would also like to thank two of my colleagues: Krishnakumar Rukmangathan for helping me with some of the diagrams for the book, and Amarpreet Singh Bassan for his help in authoring the chapters on troubleshooting. You guys were of great help. Without your input, it would have been a struggle and the book would have been incomplete.

Last but not least, I must acknowledge all the support and encouragement provided by my good friends Sneha Deep Chowdhury and Soumendu Mukherjee. Though you are experts in completely different technical domains, you guys have always been there with me listening patiently about the progress of the book, the hurdles faced and what not, from the beginning to the end. Thanks for being there with me through all my blabberings.
Introduction

My journey in Big Data started back in 2012 in one of our unit meetings. Ranjan Bhattacharjee (our boss) threw in some food for thought with his questions: “Do you guys know Big Data? What do you think about it?” That was the first time I heard the phrase “Big Data.” His inspirational speech on Big Data, Hadoop, and future trends in the industry, triggered the passion for learning something new in a few of us.

Now we are seeing results from a historic collaboration between open source and proprietary products in the form of Microsoft HDInsight. Microsoft and Apache have joined hands in an effort to make Hadoop available on Windows, and HDInsight is the result. I am a big fan of such integration. I strongly believe that the future of IT will be seen in the form of integration and collaboration opening up new dimensions in the industry.

The world of data has seen exponential growth in volume in the past couple of years. With the web integrated in each and every type of device, we are generating more digital data every two years than the volume of data generated since the dawn of civilization. Learning the techniques to store, manage, process, and most importantly, make sense of data is going to be key in the coming decade of data explosion. Apache Hadoop is already a leader as a Big Data solution framework based on Java/Linux. This book is intended for readers who want to get familiar with HDInsight, which is Microsoft’s implementation of Apache Hadoop on Windows.

Microsoft HDInsight is currently available as an Azure service. Windows Azure HDInsight Service brings in the user friendliness and ease of Windows through its blend of Infrastructure as a Service (IaaS) and Platform as a Service (PaaS). Additionally, it introduces .NET and PowerShell based job creation, submission, and monitoring frameworks for the developer communities based on Microsoft platforms.

Intended Audience

*Pro Microsoft HDInsight* is intended for people who are already familiar with Apache Hadoop and its ecosystem of projects. Readers are expected to have a basic understanding of Big Data as well as some working knowledge of present-day Business Intelligence (BI) tools. This book specifically covers HDInsight, which is Microsoft’s implementation of Hadoop on Windows. The book covers HDInsight and its tight integration with the ecosystem of other Microsoft products, like SQL Server, Excel, and various BI tools. Readers should have some understanding of those tools in order to get the most from this book.

Versions Used

It is important to understand that HDInsight is offered as an Azure service. The upgrades are pretty frequent and come in the form of Azure Service Updates. Additionally, HDInsight as a product has core dependencies on Apache Hadoop. Every change in the Apache project needs to be ported as well. Thus, you should expect that version numbers of several components will be updated and changed going forward. However, the crux of Hadoop and HDInsight is not going to change much. In other words, the core of this book’s content and methodologies are going to hold up well.
Structure of the Book

This book is best read sequentially from the beginning to the end. I have made an effort to provide the background of Microsoft’s Big Data story, HDInsight as a technology, and the Windows Azure Storage infrastructure. This book gradually takes you through a tour of HDInsight cluster creation, job submission, and monitoring, and finally ends with some troubleshooting steps.

Chapter 1 – “Introducing HDInsight” starts off the book by giving you some background on Big Data and the current market trends. This chapter has a brief overview of Apache Hadoop and its ecosystem and focuses on how HDInsight evolved as a product.

Chapter 2 – “Understanding Windows Azure HDInsight Service” introduces you to Microsoft’s Azure-based service for Apache Hadoop. This chapter discusses the Azure HDInsight service and the underlying Azure storage infrastructure it uses. This is a notable difference in Microsoft’s implementation of Hadoop on Windows Azure, because it isolates the storage and the cluster as a part of the elastic service offering. Running idle clusters only for storage purposes is no longer the reality, because with the Azure HDInsight service, you can spin up your clusters only during job submission and delete them once the jobs are done, with all your data safely retained in Azure storage.

Chapter 3 – “Provisioning Your HDInsight Service Cluster” takes you through the process of creating your Hadoop clusters on Windows Azure virtual machines. This chapter covers the Windows Azure Management portal, which offers you step-by-step wizards to manually provision your HDInsight clusters in a matter of a few clicks.

Chapter 4 – “Automating HDInsight Cluster Provisioning” introduces the Hadoop .NET SDK and Windows PowerShell cmdlets to automate cluster-creation operations. Automation is a common need for any business process. This chapter enables you to create such configurable and automatic cluster-provisioning based on C# code and PowerShell scripts.

Chapter 5 – “Submitting Jobs to Your HDInsight Cluster” shows you ways to submit MapReduce jobs to your HDInsight cluster. You can leverage the same .NET and PowerShell based framework to automate your data processing operations and retrieve the output. This chapter also teaches you how to create a MapReduce job in .NET. Again, this is unique in HDInsight, as traditional Hadoop jobs are based on Java only.

Chapter 6 – “Exploring the HDInsight Name Node” discusses the Azure virtual machine that acts as your cluster’s Name Node when you create a cluster. You can log in remotely to the Name Node and execute command-based Hadoop jobs manually. This chapter also speaks about the web applications that are available by default to monitor cluster health and job status when you install Hadoop.

Chapter 7 – “Using the Windows Azure HDInsight Emulator” introduces you to the local, one-box emulator for your Azure service. This emulator is primarily intended to be a test bed for testing or evaluating the product and your solution before you actually roll it out to Azure. You can simulate both the HDInsight cluster and Azure storage so that you can evaluate it absolutely free of cost. This chapter teaches you how to install the emulator, set the configuration options, and test run MapReduce jobs on it using the same techniques.

Chapter 8 – “Accessing HDInsight over Hive and ODBC” talks about the ODBC endpoint that the HDInsight service exposes for client applications. Once you install and configure the ODBC driver correctly, you can consume the Hive service running on HDInsight from any ODBC-compliant client application. This chapter takes you through the download, installation, and configuration of the driver to the successful connection to HDInsight.
Chapter 9 – “Consuming HDInsight from Self-Service BI Tools” is a particularly interesting chapter for readers who have a BI background. This chapter introduces some of the present-day, self-service BI tools that can be set up with HDInsight within a few clicks. With data visualization being the end goal of any data-processing framework, this chapter gets you going with creating interactive reports in just a few minutes.

Chapter 10 – “Integrating HDInsight with SQL Server Integration Services” covers the integration of HDInsight with SQL Server Integration Services (SSIS). SSIS is a component of the SQL Server BI suite and plays an important part in data-processing engines as a data extract, transform, and load tool. This chapter guides you through creating an SSIS package that moves data from Hive to SQL Server.

Chapter 11 – “Logging in HDInsight” describes the logging mechanism in HDInsight. There is built-in logging in Apache Hadoop; on top of that, HDInsight implements its own logging framework. This chapter enables readers to learn about the log files for the different services and where to look if something goes wrong.

Chapter 12 – “Troubleshooting Cluster Deployments” is about troubleshooting scenarios you might encounter during your cluster-creation process. This chapter explains the different stages of a cluster deployment and the deployment logs on the Name Node, as well as offering some tips on troubleshooting C# and PowerShell based deployment scripts.

Chapter 13 – “Troubleshooting Job Failures” explains the different ways of troubleshooting a MapReduce job-execution failure. This chapter also speaks about troubleshooting performance issues you might encounter, such as when jobs are timing out, running out of memory, or running for too long. It also covers some best-practice scenarios.

**Downloading the Code**

The author provides code to go along with the examples in this book. You can download that example code from the book's catalog page on the Apress.com website. The URL to visit is [http://www.apress.com/9781430260554](http://www.apress.com/9781430260554). Scroll about halfway down the page. Then find and click the tab labeled Source Code/Downloads.

**Contacting the Author**

You can contact the author, Debarchan Sarkar, through his twitter handle @debarchans. You can also follow his Facebook group at [https://www.facebook.com/groups/bigdatalearnings/](https://www.facebook.com/groups/bigdatalearnings/) and his Facebook page on HDInsight at [https://www.facebook.com/MicrosoftBigData](https://www.facebook.com/MicrosoftBigData).