BIG DATA
with IBM Database
ANALYSIS

ABSTRACT

ABSTRACT

Big data has emerged in the past few years as a new paradigm providing abundant data and opportunities to improve and/or enable research and decision-support applications with unprecedented value for digital earth applications including business, sciences and engineering. At the same time, Big Data presents challenges for digital earth to store, transport, process, mine and serve the data. Cloud computing provides fundamental support to address the challenges with shared computing resources including computing, storage, networking and analytical software; the application of these resources has fostered impressive Big Data advancements. This paper surveys the two frontiers –Big Data and cloud computing –and reviews the advantages and consequences of utilizing cloud computing to tackling Big Data in the digital earth and relevant science domains.

INTRODUCTION

Big Data refers to the flood of digital data from many digital earth sources, including sensors, digi-tizers, scanners, numerical modeling, mobile phones, Internet, videos, e-mails and social networks.The data types include texts, geometries, images, videos, sounds and combinations of each. Such datacan be directly or indirectly related to geospatial information . The evolution of technologies and human understanding of the data have shift datahandling from the more traditional static mode to an accelerating data arena characterized by volume, velocity, variety, veracity and value.

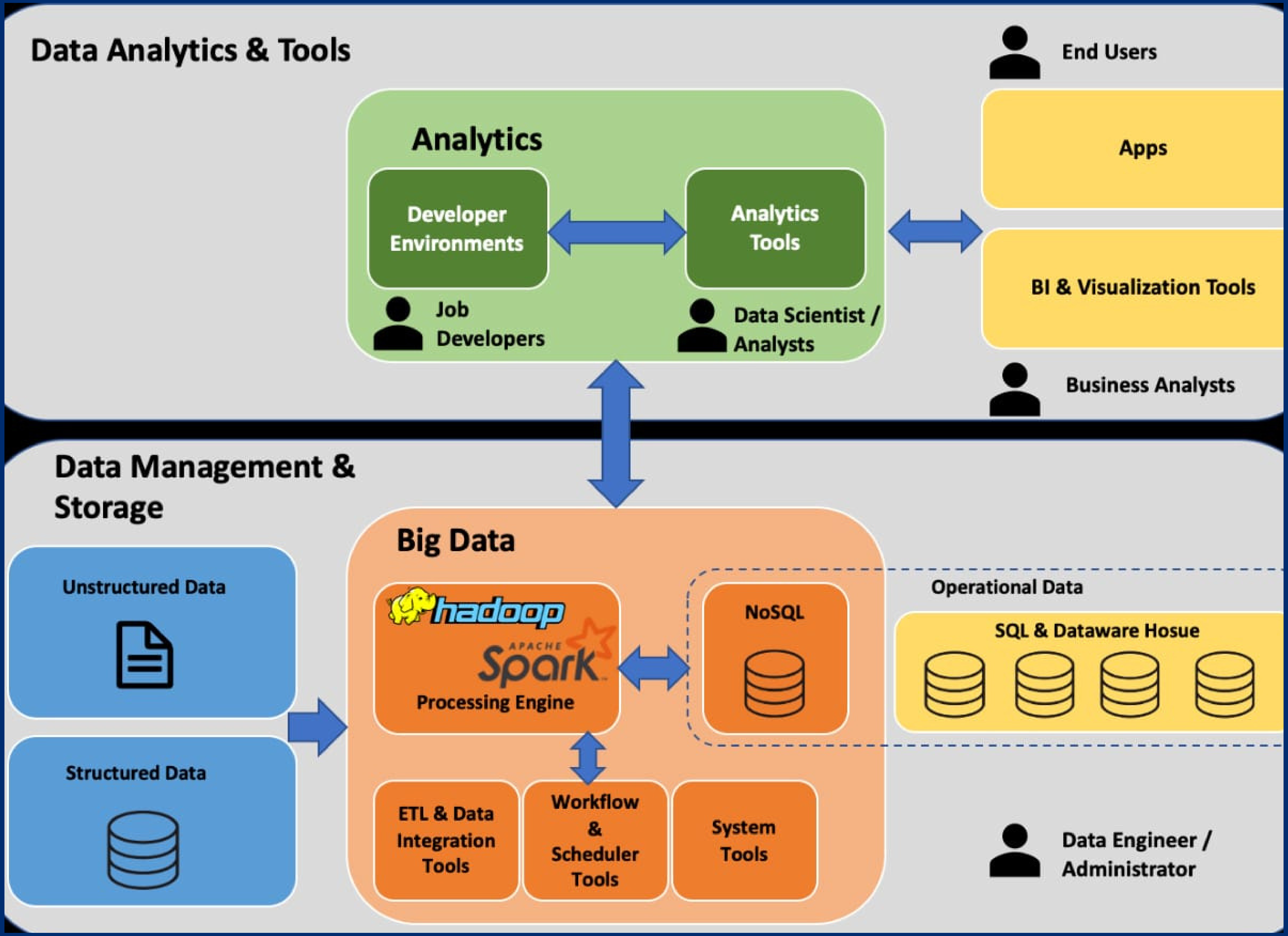
SOURCES OF BIG DATA

While Big Data has the feature of 5Vs, the feature-based challenges vary in different digital earthrelevant domains. This section reviews relevant domain-specific Big Data challenges in the sequenceof how closely are they related to geospatial principles .

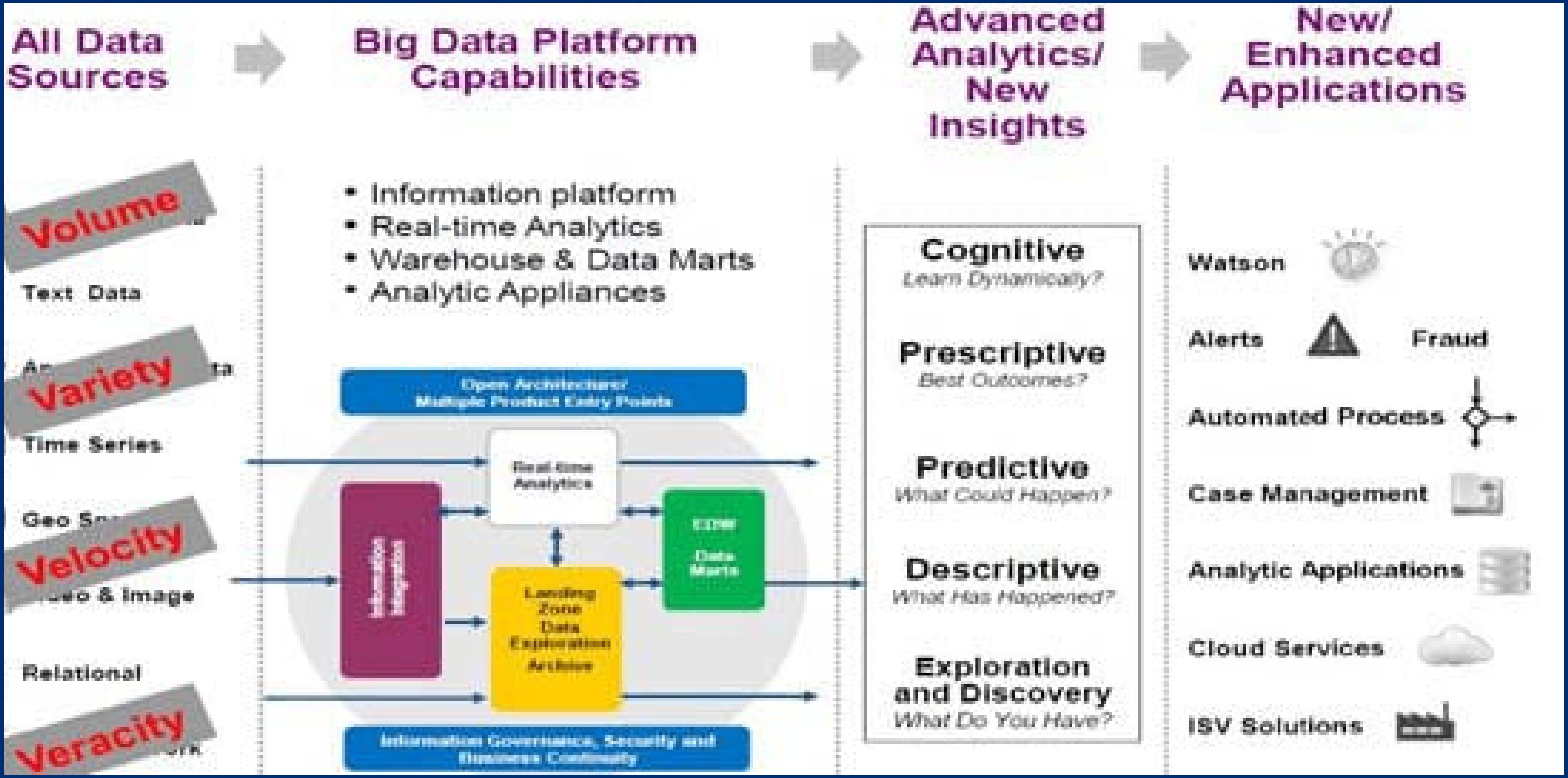
DATA STORAGE

Storage challenges are posed by the volume, velocity and variety of Big Data. Storing Big Data ontraditional physical storage is problematic as hard disk drives (HDDs) often fail, and traditionaldata protection mechanisms (e.g. RAID or redundant array of independent disks) are not efficientwith PB-scale storage (Robinson 2012). In addition, the velocity of Big Data requires the storagesystems to be able to scale up quickly which is difficult to achieve with traditional storage systems.Cloud storage services (e.g. Amazon S3, Elastic Block Store or EBS) offer virtually unlimited storagewith high fault tolerance which provides potential solutions to address Big Data storage challenges.

DATA ANALYTICS



THE 5 V'S



All Data Sources

Volume

Text Data

An

Variety

ta

Time Series

Geo Sp

Velocity

deo & Image

Relational

Veracity

rk

Big Data Platform Capabilities

- Information platform
- Real-time Analytics
- Warehouse & Data Marts
- Analytic Appliances

Open Architecture/
Multiple Product Entry Points

Real-time Analytics

Information Integration

Landing Zone
Data Exploration
Archive

EDW

Data Marts

Information Governance, Security and Business Continuity

Advanced Analytics/ New Insights

Cognitive
Learn Dynamically?

Prescriptive
Best Outcomes?

Predictive
What Could Happen?

Descriptive
What Has Happened?

Exploration and Discovery
What Do You Have?

New/ Enhanced Applications

Watson

Alerts

Fraud

Automated Process

Case Management

Analytic Applications

Cloud Services

ISV Solutions

DATA SECURITY

The increasing dependence on computers and Internet over the past decades makes businesses and individuals vulnerable to data breach and abuse. Big Data poses new security challenges for traditional data encryption standards, methodologies and algoritm.Previous studies of data encryption focused on small-to-medium-size data,which does not work well for Big Data due to issues of the performance and scalability.

DATA QUALITY

Data quality includes four aspects: accuracy, completeness, redundancy and consistency . The intrinsic nature of complexity and heterogeneity of Big Data makes data accuracy andcompleteness difficult to identify and track, thus increasing the risk of ‘false discoveries’.For example, social media data are highly skewed in space, time and demographics, and locationaccuracy varies from meters to hundreds of kilometers

CURRENT STATUS OF TACKLING BIG DATA CHALLENGES

While the Big Data challenges can be tackled by many advanced technologies, such as HPC, cloudcomputing is the most elusive and important.

CONCLUSION

The 5Vs that characterize Big Data and the five features of cloud computing increasingly play adominant role in this innovation process for digital earth. The current innovation opportunities andresearch agenda of utilizing cloud computing for tackling Big Data are summarized.