










海量数据分成两块，一是系统建设技术，二，海量数据应用。

先说系统建设，现在主流的技术是HADOOP，主要基于mapreduce的分布式框架。目前可以先学习这个。但是我的观点，在分布式系统出来之前，主要是集中式架构，如DB2，oracle。为什么现在用分布式架构，那是因为现在集中式架构受限于IO性能，出来速度慢，如果又一种硬件技术，可以很快地处理海量数据，性能上能满足需求，那么集中式架构优于分布式架构，因为集中式架构稳定，运维压力小。现在的集中式架构要么性能达不到要求，要么就是过于昂贵。我期待一种技术出现，可以非常快地传输和处理数据，那么集中式架构将再次进入人们眼球。再说海量数据应用。海量数据应用主要是数据挖掘和机器算法。具体有不同的应用场景，如个性化搜索和推荐，社交网络发现，精准营销，精准广告，实时最优路径，人工智能等等。看你想做系统支撑技术还是与业务结合的应用技术。

如果现在学系统建设技术，可以读下如下书籍：

-  [Getting.Started.with.CouchDB(2012.01)].MC.Brown.文字版
-  [Google.Script : Enterprise.Application.Essentials(2012.01)].James.Ferreira.文字版
-  [Hadoop : The.Definitive.Guide(3rd,Early.Release)].Tom.White.文字版
-  [Learning.Pperl.Student.Workbook(2nd,2012.01)].brian.d.foy.文字版
-  [MongoDB.and.PHP(2012.01)].Steve.Francia.文字版
-  [Node.for.Front-End.Developers(2012.01)].Garann.Means.文字版
-  [SQL.in.a.Nutshell(3rd,2008.11)].Kevin.Kline.文字版
-  [Web.2.0 : A.Strategy.Guide(2008.04)].Amy.Shuen.文字版
-  [Writing.and.Querying.MapReduce.Views.in.CouchDB(2011.01)].Bradley.Holt.文字版

系统建设技术书籍

如果学数据挖掘和机器算法，推荐先看数据挖掘导论，统计分析原理，Mahout，R，MATLAB