

<https://youtu.be/QDBeUo3k0Nc?si=msT4VcQf4-rtzWrr>

<https://youtu.be/xgPa7RGOyGw?si=cGD2t6idMandH4Nm>

<https://towardsdatascience.com/hits-algorithm-link-analysis-explanation-and-python-implementation-61f0762fd7cf>

The HITS (Hyperlink-Induced Topic Search) algorithm is a link analysis algorithm used to rank web pages based on their authority and hub scores. It is used in web search engines to determine the importance of web pages within a network of interconnected pages. HITS is not limited to big data analytics but can be adapted for such applications. Here's an overview of how the HITS algorithm works and how it can be applied in the context of big data analytics:

### **1. Authority and Hub Scores:**

- In the context of HITS, each web page is assigned two scores: authority and hub scores.
- Authority score: A measure of a web page's importance based on the number of quality inbound links it receives.
- Hub score: A measure of a web page's ability to link to other authoritative pages.

### **2. Iterative Algorithm:**

- The HITS algorithm is iterative and involves updating the authority and hub scores of web pages until convergence is achieved. Typically, it's implemented using the power iteration method.
- At the beginning, all web pages are given equal authority and hub scores.

### **3. Link Analysis:**

- During each iteration, the algorithm analyzes the links between pages to compute updated authority and hub scores.
- Pages that link to authoritative pages receive higher hub scores.
- Pages that are linked to by authoritative pages receive higher authority scores.

### **4. Convergence:**

- The algorithm continues to iterate until the scores of the web pages stabilize or reach a predefined convergence criteria (e.g., a specific number of iterations or small changes in scores).

### **5. Application in Big Data Analytics:**

- In big data analytics, the HITS algorithm can be applied to various scenarios, not just web pages.
- For example, in a big data context, you might have a network of data entities, and you want to identify the most authoritative and influential entities in the network.
- This can be used in applications such as social network analysis, citation analysis, recommendation systems, and more.

- You can adapt the HITS algorithm to analyze large datasets and networks by using distributed computing frameworks like Apache Hadoop or Apache Spark.